

State-of-the-Art  
Survey

Yang Cai  
Julio Abascal (Eds.)

LNAI 3864

# Ambient Intelligence in Everyday Life

Foreword by Emile Aarts



 Springer

Lecture Notes in Artificial Intelligence 3864

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Yang Cai Julio Abascal (Eds.)

# Ambient Intelligence in Everyday Life

Foreword by Emile Aarts

## Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

## Volume Editors

Yang Cai

Carnegie Mellon University  
Ambient Intelligence Studio, Cylab  
CIC 2218, 4720 Forbes Avenue, Pittsburgh, PA 15213, USA  
E-mail: ycai@cmu.edu

Julio Abascal

University of the Basque Country  
Dept. of Computer Architecture & Technology, School of Informatics  
Laboratory of Human-Computer Interaction for Special Needs  
Manuel Lardizabal 1, 20018 Donostia, Spain  
E-mail: julio.abascal@si.ehu.es

Library of Congress Control Number: 2006931064

CR Subject Classification (1998): I.2, H.5, H.4, C.2, D.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743  
ISBN-10 3-540-37785-9 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-37785-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11825890 06/3142 5 4 3 2 1 0



An ordinary summer night in Donostia-San Sebastian, Spain, where Ambient Intelligence for Everyday Life Workshop was held on July 21-22, 2005.

# Foreword

Back in 1997, on the occasion of the 50th anniversary of the Association of Computing Machinery, computer scientists from all over the world were asked for their opinion about the next 50 years of computing. Although rooted in many different disciplines, the scientists' reactions were strikingly consistent in the sense that they all envisioned a world consisting of distributed computing devices that would surround people in a non-obtrusive way. As one of the first paradigms following this vision, Marc Weiser's Ubiquitous Computing was aimed at a novel computing infrastructure that would replace the current mobile computing infrastructure by interconnected, transparent, and embedded devices that would facilitate ubiquitous access to any source of information at any place, any point in time and by any person. Such a world could be conceived by a huge distributed network consisting of thousands of interconnected embedded systems surrounding the user and satisfying his or her needs for information, communication, navigation, and entertainment.

Ambient Intelligence (AmI), introduced in the late 1990s as a novel paradigm for digital systems for the years 2010-2020, builds on the early ideas of Weiser by taking the embedding and integration one step further. This disruptive improvement may be conceived by embedding computational intelligence into the networked environment and moving the user into the foreground by supporting him or her with intuitive and natural interaction concepts. According to the definition, ambient intelligence refers to smart electronic environments that are sensitive and responsive to the presence of people. Since its introduction, this vision has grown mature, having become quite influential in the development of new concepts for information processing as well as combining multidisciplinary fields including electrical engineering, computer science, industrial design, user interfaces and cognitive sciences.

The AmI paradigm provides a basis for new models of technological innovation within a multidimensional society. The essential enabling factor of the AmI vision is provided by the fact that current technological developments indeed enable the large-scale integration of electronics into the environment, thus enabling the actors, i.e., people and objects to interact with their environment in a seamless, trustworthy and natural manner. In addition, the past years reveal a growing interest in the role of information and communication technology to support peoples' lives, not only for the purpose of increased productivity, but also for the purpose of self-expression, health-care, leisure, and creativity. A major issue in this respect is given by the growing awareness that novel products such as devices and services should meet elementary user requirements, i.e., usefulness and simplicity. Hence, it is generally believed that novel technologies should not increase functional complexity, but should merely contribute to the development of *easy to use* and *simple to experience* products. Obviously, this statement has a broad endorsement by a wide community of both designers and engineers, but reality reveals that it is hard to achieve in practice, and

that novel approaches, as may be provided by the AmI vision, are needed to make it work.

All these new socio-economic developments open up major opportunities for making money in markets that exploit ambient intelligent technology, thus providing the necessary economical foundation for the development of ambient intelligence. The developments in ambient intelligence obtained during its relative short existence reveal that the vision is rapidly gaining traction, and that some of the early ideas following from this vision are growing mature. Examples are the use of context-aware services in mobile phones exploiting RFID tags and personalized media players that apply user profiles and collaborative filtering techniques to recommend and exchange digital information. No matter how smart and technologically advanced these examples may be, they cannot be regarded as the major and the most convincing breakthroughs in ambient intelligence that are needed to lead to full acceptance of the new paradigm.

As the success of the AmI paradigm relies heavily on the social acceptance of the newly proposed ambient technology, we need to look at the human factors' side of the vision in order to study the relation between AmI technology and people's behavior. This would reveal the true added value of ambient intelligence in everyday life. In view of this, the present special LNAI volume on *Ambient Intelligence in Everyday Life* can be considered a timely document that provides an important contribution to the dissemination of the AmI vision. The editors have succeeded in bringing together a quite interesting and inspiring collection of research contributions reporting on the progress of ambient intelligence within various domains. Divided in three parts, i.e., "Human Centered Computing," "Ambient Interfaces," and "Architectures in Ambient Intelligence," the volume presents a fine collection of papers emphasizing the multi-disciplinary character of the investigations not only from the point of view of the various scientific fields involved, but also from an application point of view. To the best of my knowledge, this is the first book that presents achievements and findings of AmI research covering the full spectrum of societally relevant applications, ranging from learning, leisure, and trust, all the way to well-being, healthcare and support for elderly and disabled persons. These make the volume truly unique, and in more than one respect, a truly valuable source of information that may be considered a landmark in the progress of ambient intelligence. Congratulations to all those who have contributed!

Eindhoven, November 2005

Emile Aarts  
Vice President of Philips Research

## Preface

As sensors and antennas are embeddable in things around us, a new era of physical Internet begins. The revolution is brought by economies of scale and millions of consumers. Today, a music birthday card may have more computing power than a mainframe computer a few decades ago. However, we still don't know how to sense human feelings electronically. Perhaps we need to go back to the drawing board and rethink our daily life.

In this volume, we focus on the cognitive aspects of ambient intelligence. In a broad sense, ambient intelligence is *perceptual interaction*, which involves sensory fusion, common sense, insight, anticipation, esthetics and emotion that we normally take for granted. We interact with the world through the windows of our senses: sight, sound, smell, taste and touch, which not only describe the nature of physical reality, but also connect us to it emotionally. Our knowledge is composed by the fusion of multidimensional information sources: shape, color, time, distance, direction, balance, speed, force, similarity, likelihood, intent and truth. Ambient intelligence is not only a perception, but also an interaction. We not only acquire information, but also construct and share information.

Common sense has been an immense challenge to ambient intelligence. For over 20 years, with over a 20-million-dollar investment, Douglas Lenat and his colleagues have been developing Cyc, a project that aims to create a reusable general knowledge base for intelligent assistants. Cyc essentially looks for a representation model of human consensual knowledge that can construct a *semantic web* where meaning is made explicit, allowing computers to process intelligently. One remarkable extension of the knowledge formalism in Cyc is the ability to handle default reasoning. In many cases, ambient intelligence operates at a default level or below perceptual thresholds. Default reasoning is generally common sense itself.

Empathic computing, or human-centric computing has been a rapidly growing area because we want an intelligent system to know the "who, what, when, and where" as it encounters different situations. In this volume, Maja Pantic reviews the-state-of-the-art face interfaces, especially in the facial emotion detection area. Furthermore, empathic computing systems such as eWatch and wearable sensor networks are explored for transforming multimodal signals into recognizable patterns. As information appliances enter our daily life, sensor-rich computing is necessary to interpret the data.

Innovative ambient interfaces are also presented in this volume, which includes co-creation in ambient narrative, hyper-reality, a pre-communication system and an ambient browser. As sound interfaces are gaining momentum because of their ambient nature, in this volume, we include papers about whistling to a mobile robot, online music search by tapping, and user-specific acoustic phased arrays.



Finally, ambient infrastructures are presented to seamlessly connect the dots between homes, offices and individuals, including a middleware structure, interfaces for elderly people at home, a smart wheelchair, and online collaborators.

The volume of work comes from the Workshop of Ambient Intelligence in Everyday Life that was held at the Miramar Congress Center, a historical building in Donostia-San Sebastian, Spain, July 21-22, 2005. The idea of organizing this small workshop was inspired by Norbert Wiener's vision about scientific communication. In his book on Cybernetics 50 years ago, he said, "the idea has been to get together a group of modest size, not exceeding some twenty in number of workers in various related fields, and to hold them together for two successive days in all-day series of informal papers, discussions, and meals together, until they have had the opportunity to thresh out their differences and to make progress in thinking along the same lines." Although the workshop was proven successful in the exchange of the different opinions, not every participant thought along the same lines about how to define ambient intelligence. This indeed reflects the nature of this dynamic field.

We are deeply in debt to all the authors and reviewers who contributed to this special volume. Without their effective support and commitment, there wouldn't be this meaningful product at all. We thank Emile Aarts from Philips Research for the Foreword. Special thanks to local workshop organizers Elena Lazkano and Basilio Sierra, the Program Committee, and all of those who were involved in the refereeing process, and all of those who helped to convene this successful workshop. We acknowledge those who have supported the research in ambient intelligence for the past years: Counselor Paul op den Brouw and Attaché Roger Kleinenberg for Science and Technology of the Royal Netherlands Embassy in Washington, DC, Program Managers Pierre-Louis Xech and Marco Combetto from Microsoft Research, Cambridge, Program Managers Karen Moe, Steven Smith, Horace Mitchell, Kai Dee Chu, Yongxiang Hu and Bin Lin from NASA, and President Karen Wolk Feinstein and Senior Program Officer Nancy Zionts from Jewish Healthcare Foundation, and colleagues from Carnegie Mellon University: Pareedp Khosla, Mel Seigel, David Kaufer, Michael Reiter, Don McGillen, Bill Eddy, Howard Watlar, Daniel Siewiorek and Raj Rajkummar.

This project was in part sponsored by grants from the University of the Basque Country-Euskal Herriko Unibertsitatea, Kutxa, Universidad del Pais Vasco, Gipuzkoako Feru Aldundia Diputacion Foral de Gipuzkoa, Robotiker, Ikerlan, NASA ESTO-AIST Program, NASA LaRC C&I Initiative and ARO.

Yang Cai and Julio Abascal  
Editors

# Organization

## Program Committee

Julio Abascal, University of the Basque Country, Spain (Co-chair)

Yang Cai, Carnegie Mellon University, USA (Co-chair)

Judith Devaney Terrill, NIST, USA

Yongxiang Hu, NASA, USA

Judith Klein-Seetharaman, Carnegie Mellon, USA

Elena Lazkano, University of the Basque Country, Spain

Ramón López de Mántaras, CSIC, Spain

Basilio Sierra, University of the Basque Country, Spain

Elena Zudilove, University of Amsterdam, The Netherlands

## Editorial Board

Julio Abascal, University of the Basque Country, Spain

Yang Cai, Carnegie Mellon University, USA

Elena Lazkano, University of the Basque Country, Spain

Basilio Sierra, University of the Basque Country, Spain

## Editorial Assistants

Nathaniel Bauernfeind, Carnegie Mellon University, USA

Mona Roxana Botezatu, Carnegie Mellon University, USA

# Table of Contents

## Part I: Human-Centric Computing

|  |    |
|--|----|
| Common Sense Reasoning – From Cyc to Intelligent Assistant . . . . .   | 1  |
| <i>Kathy Panton, Cynthia Matuszek, Douglas Lenat, Dave Schneider,<br/>Michael Witbrock, Nick Siegel, Blake Shepard</i> |    |
| Face for Ambient Interface . . . . .   | 32 |
| <i>Maja Pantic</i>   |    |
| Empathic Computing . . . . .   | 67 |
| <i>Yang Cai</i>  |    |
| Location and Activity Recognition Using eWatch: A Wearable Sensor<br>Platform . . . . .                                | 86 |
| <i>Uwe Maurer, Anthony Rowe, Asim Smailagic, Daniel Siewiorek</i>  |    |

## Part II: Ambient Interfaces

|   |     |
|---|-----|
| Co-Creation in Ambient Narratives . . . . .   | 103 |
| <i>Mark van Doorn, Arjen P. de Vries</i>  |     |
| Living with Hyper-reality . . . . .   | 130 |
| <i>Leonardo Bonanni</i>   |     |
| Ambient Pre-Communication . . . . .   | 142 |
| <i>Atsunobu Kimura, Yoshihiro Shimada, Minoru Kobayashi</i>   |     |
| AmbientBrowser: Web Browser in Everyday Life . . . . .  | 157 |
| <i>Satoshi Nakamura, Mitsuru Minakuchi, Katsumi Tanaka</i>  |     |
| Online Music Search by Tapping . . . . .  | 178 |
| <i>Geoffrey Peters, Diana Cukierman, Caroline Anthony,<br/>Michael Schwartz</i>                       |     |
| Whistling to Machines . . . . .   | 198 |
| <i>Urko Esnaola, Tim Smithers</i>   |     |
| Speaker Identification and Speech Recognition Using Phased Arrays . . . . .                           | 227 |
| <i>Roger Xu, Gang Mei, ZuBing Ren, Chiman Kwan, Julien Aube,<br/>Cedrick Rochet, Vincent Stanford</i> |     |

### Part III: Architectures in Ambient Intelligence

|  |            |
|--|------------|
| A Middleware for the Deployment of Ambient Intelligent Spaces . . . . .  | 239        |
| <i>Diego López-de-Ipiña, Juan Ignacio Vázquez, Daniel García,<br/>Javier Fernández, Iván García, David Sáinz, Aitor Almeida</i>              |            |
| Ambient Interfaces for Elderly People at Home . . . . .  | 256        |
| <i>Fausto J. Sainz Salces, Michael Baskett, David Llewellyn-Jones,<br/>David England</i>   |            |
| A Smart Electric Wheelchair Using UPnP . . . . .   | 285        |
| <i>Daniel Cascado, Saturnino Vicente, J. Luis Sevillano,<br/>Claudio Amaya, Alejandro Linares, Gabriel Jiménez,<br/>Antón Civit-Balcells</i> |            |
| Collaborative Discovery Through Biological Language Modeling<br>Interface . . . . .  | 300        |
| <i>Madhavi Ganapathiraju, Vijayalaxmi Manoharan, Raj Reddy,<br/>Judith Klein-Seetharaman</i>   |            |
| <b>Author Index . . . . .</b>  | <b>323</b> |

# Common Sense Reasoning – From Cyc to Intelligent Assistant

Kathy Panton, Cynthia Matuszek, Douglas Lenat, Dave Schneider,  
Michael Witbrock, Nick Siegel, and Blake Shepard

Cycorp, Inc.

3721 Executive Center Drive, Suite 100, Austin, TX 78731, USA

{panton, cyndy, lenat, daves, witbrock, nsiegel, blake}@cyc.com

**Abstract.** Semi-formally represented knowledge, such as the use of standardized keywords, is a traditional and valuable mechanism for helping people to access information. Extending that mechanism to include formally represented knowledge (based on a shared ontology) presents a more effective way of sharing large bodies of knowledge between groups; reasoning systems that draw on that knowledge are the logical counterparts to tools that perform well on a single, rigidly defined task. The underlying philosophy of the Cyc Project is that software will never reach its full potential until it can react flexibly to a variety of challenges. Furthermore, systems should not only handle tasks automatically, but also actively anticipate the need to perform them. A system that rests on a large, general-purpose knowledge base can potentially manage tasks that require world knowledge, or “common sense” – the knowledge that every person assumes his neighbors also possess. Until that knowledge is fully represented and integrated, tools will continue to be, at best, *idiots savants*. Accordingly, this paper will in part present progress made in the overall Cyc Project during its twenty-year lifespan – its vision, its achievements thus far, and the work that remains to be done. We will also describe how these capabilities can be brought together into a useful ambient assistant application.

Ultimately, intelligent software assistants should dramatically reduce the time and cognitive effort spent on infrastructure tasks. Software assistants should be *ambient systems* – a user works within an environment in which agents are actively trying to classify the user’s activities, predict useful subtasks and expected future tasks, and, proactively, perform those tasks or at least the sub-tasks that can be performed automatically. This in turn requires a variety of necessary technologies (including script and plan recognition, abductive reasoning, integration of external knowledge sources, facilitating appropriate knowledge entry and hypothesis formation), which must be integrated into the Cyc reasoning system and Knowledge Base to be fully effective.

## 1 The Evolution of Cyc

### 1.1 Beginnings of the Cyc Project

In the early 1970s, rule-based expert systems such as MYCIN [18] and DENDRAL [3] were AI’s major success story. MYCIN acted as an assistant in the diagnosis of

blood infections, while DENDRAL's expertise was in chemical analysis. These applications used rules to solve problems within circumscribed domains. Expert systems represented a major step forward in AI technology and are used today to address problems as diverse as camera lens design and cargo placement [5], but their limitations quickly became obvious. Lenat and Guha [10] provide several examples of the *brittleness* displayed by expert systems. Two will suffice here:

An expert system authorizes a car loan to someone who stated, on his application, that he'd worked at the same job for twenty years. A good risk? Perhaps, but the individual also stated he was 18 years old.

A skin disease diagnosis program is told about a "patient" that is a 1969 Chevrolet:

**Program:** Are there spots on the body?

**User:** Yes.

**Program:** What color spots?

**User:** Reddish-brown.

**Program:** Are there more spots on the trunk than elsewhere?

**User:** No.

**Program:** The patient has measles.

In the first example, the system failed to notice what was likely to have been a simple typo; perhaps the applicant meant that he had been at his current job for two years, or 20 months. Rules encoded in that system might nevertheless conclude that someone employed by the same company for 20 years is a very good credit risk, resulting in an easy loan approval. The system breaks down because it cannot detect what, to humans, are very obvious contradictions. These errors can have effects far more dire than in the car loan case. For example, a medical transcriptionist accidentally transposing a patient's weight and age in a patient record could lead to that patient being prescribed medications at dangerously incorrect dosage levels.

The second example illustrates that expert systems work only within the domain for which they were explicitly engineered; this software cannot correctly diagnose rust spots on a car. Furthermore, the system is unable to use its knowledge about skin infections to do things like recommend treatment or explain to users how long the disease might last and what other symptoms the patient may be experiencing. In short, this software contains many handcrafted rules that encode useful information about skin diseases, but this knowledge is isolated and opaque: it is useless when applied to an object outside its domain, and cannot be reused across similar or related problems.

Expert systems have no understanding of what they are for, or the extent of their own knowledge. But their brittleness is mainly due to a lack of *common sense*. This is the general knowledge that allows us to get by in the real world, and to flexibly understand and react to novel situations. We remember and use (though usually not consciously) heuristics such as "Water makes things wet"; "Wet metal may rust"; "No two objects can occupy the same space at the same time"; and "Inanimate objects don't get diseases".

The driving force behind the Cyc Project was the realization that almost all software programs would benefit from the application of common sense. Expert systems would gain protection against user error or intentional fraud; the consistency of data in spreadsheets could be checked automatically; information-retrieval systems and word processors could exhibit more useful behaviors based on an understanding of the user's goals at any given point. The greatest impediment to the achievement of AI was the inability of programs to accumulate, apply, and reuse general knowledge.

Lenat began the Cyc Project in 1984, at the Microelectronics and Computer Technology Corporation in Austin, Texas, with the goal of building a single intelligent agent. This agent would be equipped not just with static facts, but also with heuristics and other problem-solving methods that would allow it to act as a substrate, an almost invisible performance-boosting layer, underlying a variety of software applications. The Cyc Project was initially envisioned as a series of ten-year, two-to-ten-person-century efforts, in (1) knowledge base and ontology building, or “pump priming”; (2) natural language understanding and interactive dialogue; and (3) automated discovery.

## 1.2 Representing Knowledge

Three preliminary research questions presented themselves: How much does a system need to know in order to be useful? What kinds of knowledge are necessary? How should this knowledge be represented?

### 1.2.1 Amount of Knowledge

The “annoying, inelegant, but apparently true” answer to the first question was that vast amounts of commonsense knowledge, representing human consensus reality, would need to be encoded to produce a general AI system (Lenat and Guha 1990) [10]. In order to mimic human reasoning, Cyc would require background knowledge regarding science, society and culture, climate and weather, money and financial systems, health care, history, politics, and many other domains of human experience. The Cyc Project team expected to encode at least a million facts spanning these and many other topic areas.

### 1.2.2 Kinds of Knowledge

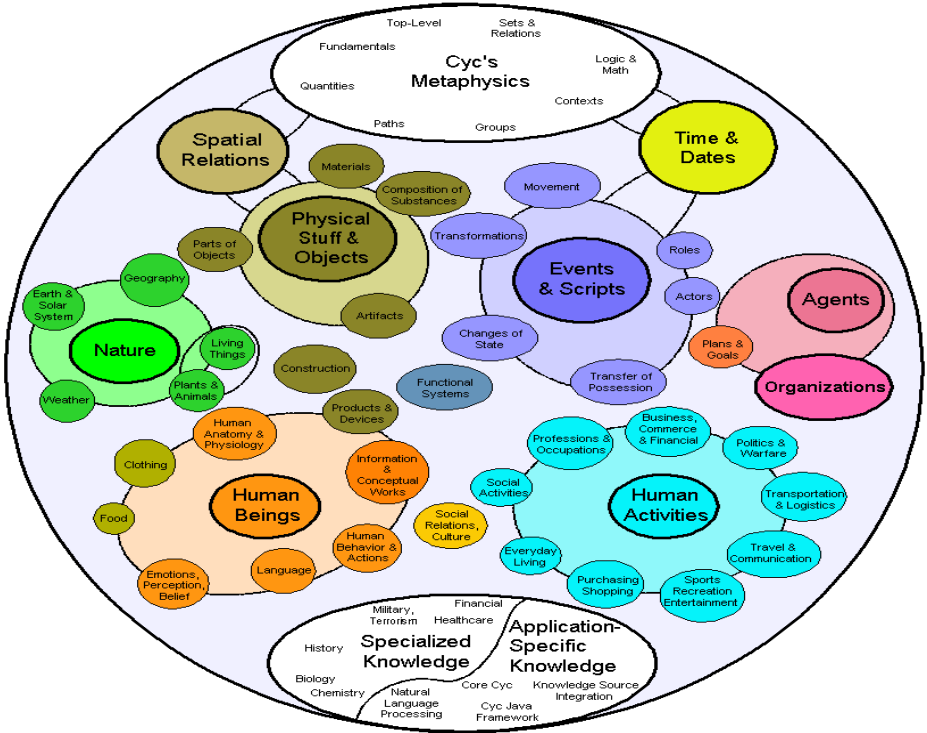
Lenat and his team understood that the “pump priming” information was not specific facts (e.g. “President Lincoln was assassinated”), but rather the “white space” – the unstated knowledge which the writer of such sentences assumes the reader already knows, such as the fact that being President requires one to be alive.

In order to truly comprehend a sentence such as “President Abraham Lincoln was assassinated”, and be able to make inferences about its likely consequences, a person must have already learned many facts about the world. Someone unable to answer the following questions cannot be said to have fully understood the example sentence:

What is a President?

Was Lincoln President two months after he was assassinated?

Was Lincoln alive 300 years before he was assassinated?



**Fig. 1.** Cyc KB Topic Map. The information in the Cyc KB can be subdivided into loosely grouped, inter-related “blocks” of knowledge at various levels of generality. In this diagram, Cyc’s understanding of metaphysics is the most general block of knowledge, gradating down to the very specific knowledge in tightly defined domains.

Can a chair be assassinated? Can a cat? An idea?

Was Lincoln’s assassin on the same continent as Lincoln when the assassination occurred?

The task, then, was transformed into one of entering not raw facts but rather a kind of pre-factual knowledge, the type of information humans learn effortlessly but rarely need to articulate.

The question of how to get this knowledge into Cyc provoked much discussion. Many AI researchers were turning to techniques such as Natural Language Understanding (NLU) or Machine Learning, in an attempt to glean knowledge from textual sources or from a seedling knowledge base. After having worked in these areas in the 1970s, however, Lenat became convinced that there was no “free lunch” in building a real intelligence. There was the chicken-and-egg problem: in order to translate text into a useful semantic representation, an NLU system needs common sense. Imagine a system with absolutely no knowledge of the world trying to determine what the ambiguous word “bank” means in the sentence, “The bank is



closed on Sundays”, or deciding to whom the pronoun “they” refers in the following sentences:

The police arrested the protestors because they feared violence.

The police arrested the protestors because they advocated violence.

It became apparent by 1983 (Lenat et al. 1983) [9] that the kind of knowledge an intelligent system must have – pre-factual consensus knowledge – was not likely to be found in any textbook, almanac or manual. Cyc could not directly learn facts such as “Once people die, they stay dead” from text sources, even if an NLU system that could handle such input were available. Cyc’s developers realized that NLU and Machine Learning would be worthwhile approaches to building the Cyc knowledge base (KB) only after a broad foundation of common sense knowledge already existed. An intelligent system (like a person) learns at the fringes of what it already knows. It follows, therefore, that the more a system knows, the more (and more easily) it can learn new facts related to its existing knowledge. It was determined that common sense must come first, and that initially it would have to be codified and entered manually into Cyc. This was the basis for the creation of the Cyc Project at MCC in Austin, in 1984.

### 1.2.3 Knowledge Representation Formalism

Now that the question of what to teach Cyc had been addressed, the next step was to find an adequate formalism – a representation language – in which to encode that knowledge. The earliest approach adopted was the now-familiar frame-and-slot language, in which terms are related by binary predicates:

GeorgeWBush

---

likesAsFriend: VladimirPutin

“George Bush considers Vladimir Putin a friend.”

Mercury

---

genls: UnalloyedMetal

“Mercury is an unalloyed metal.”

However, limitations of the frame-and-slot method quickly became apparent. First, it was impossible to make quantified statements such as “All mammals have legs.” Second, modal operators such as “believes” and “desires”, which require embedded clauses, could not be expressed; nor could other (often implicit) aspects of context such as time, space, and number be captured (e.g., “In modern Western cultures, adults are generally permitted to have only one spouse at a time.”).

Clearly, more expressivity was needed. Other desiderata for a representation language fell out from the requirement that Cyc be able to perform useful types of reasoning over its knowledge. In summary, Cyc’s representation formalism needed to:

- Have a clear, simple, declarative semantics;
- Include conjunctions, disjunctions, quantifiers, equality, and inequality operators;

- Allow for meta-level assertions, or statements about statements (e.g., “This rule was entered by Doug Lenat on March 2, 1986”);
- Support inference mechanisms such as verifying conjectures, finding appropriate bindings for variables, and planning;
- Allow nested expressions, such as those found in statements of propositional attitudes (e.g., “Joe believes that Tom’s birthday is tomorrow.”).

### 1.3 CycL: Cyc’s Representation Language

Cyc’s representation language is known as CycL. It is essentially an augmented version of first-order predicate calculus (FOPC). All of the FOPC connectives, such as *and*, *or*, and *implies*, are present, as are the quantifiers. One crucial extension was the ability to handle *default reasoning*; aside from intrinsically definitional information (e.g., “All dogs are mammals”), there are few general statements one can make about the world that cannot have exceptions. Some examples:

You can see other peoples’ noses, but not their hearts.

Given two professions, either one is a specialization of the other, or else they are likely to be practiced independently.

Dogs have four legs.

These statements are usually true, but three-legged dogs do exist; thoracic surgeons routinely see their patients’ hearts; and there are some people who practice two separate professions simultaneously. Therefore, standard truth-conditional logic, in which statements are only either true or false, would not suffice. Currently, every assertion in the KB carries a truth value that indicates its degree of truth. CycL contains five possible truth values: *monotonically false*, *default false*, *unknown*, *default true*, and *monotonically true*. Most assertions in the KB are either *default true* or *monotonically true*. Default assertions can be overridden by new knowledge, whether it comes from a person using Cyc or is derived by Cyc’s own inference engine. Instead of using only a single support or line of reasoning to determine if an assertion is true or false, Cyc’s inference engine uses argumentation. This is the process of weighing various arguments, pro and con, to arrive at a truth value for the assertion. Cyc employs a number of heuristics during argumentation. One simple example is to prefer monotonic rules: if two rules conclude P but with different truth values (i.e., one concludes that P is monotonically true but the other concludes that P is default false), then, all else being equal, Cyc sets the truth value of P to the one suggested by the monotonic rule.

Arguments consist of justification chains showing which proofs (ground facts, rules, and inference methods) were used to arrive at a conclusion. Figure 2 shows a partial inference chain constructed in response to the query, “Can the Earth run a marathon?”

### 1.4 Structure of the Cyc KB

The Cyc KB consists of terms representing individuals (e.g., *CityOfParisFrance*, *BillClinton*) and natural kinds (*Platinum*, *PineTree*). Cyc predicates can

have as many arguments as are appropriate, though one-, two-, and three-place predicates are most common. Functions that can be applied to existing terms in order to construct new ones – e.g., `LiquidFn` and `BorderBetweenFn` – permit the compositional reification of an unlimited number of new “non-atomic” collections and individuals, such as `(LiquidFn Nitrogen)` and `(BorderBetweenFn Sweden Norway)`.

Early on, it became clear that a very large knowledge base would eventually run into the problem of internal inconsistency: statements would begin to contradict one another. Such contradictions do not necessarily indicate a fault in the formulation of the assertions themselves; humans encounter this phenomenon on a regular basis. Imagine the following conversation:

CHILD: Who is Dracula, Dad?  
FATHER: A vampire.  
CHILD: Are there really vampires?  
FATHER: No, vampires don't exist.

This exchange is contradictory at first blush, but does not remain so when we consider that the truth of each statement depends on some implicit *context of reference*. In answering the child's first question, the father frames his response within the context of mythology and fiction. His second answer, however, is framed in a real-world context. In the fictional world of vampires, it is true that they exist, tend to be pale, and have sharp canine teeth used for puncturing necks. In the actual world, there are no vampires (though there are vampire stories).

Since it is impossible to maintain global consistency in a knowledge base containing millions of assertions, the Cyc Project's approach was to aim for local consistency instead; this approach was later extended in the thesis work of R.V. Guha [10]. This is achieved by situating each assertion in a particular *microtheory* (essentially, an explicitly represented logical context). Assertions within a microtheory must be consistent with one another, but need not be consistent with those in other microtheories. Additionally, bundles of assertions in a microtheory tend to share many background assumptions. The microtheory mechanism allows these assumptions to be stated once, at the level of the microtheory, instead of having to be repeated for each affected assertion. For example, the microtheory named `NormalPhysicalConditionsMt` contains assertions such as “Fluorine is a gas” and “Glass has a high amount of shear strength”. These assertions share the domain assumptions that temperature, pressure, etc. are in the normal range for Earth's surface.

Cyc's microtheory mechanism also allows for more efficient inference, in many cases. When solving certain types of problems, Cyc knows, or is told, that some microtheories must be consulted, while others are irrelevant. This reduces the search space, speeding up the inference process.

## 1.5 Addressing Structured External Knowledge in Cyc

In the Cyc KB, it is useful to make the distinction between *knowledge* (the underlying heuristics that allow us to reason) and *data* (facts or statements about specific items in

```
Type : TRANS-PREDICATE-NEGATIONPREDS-NEG
```

```
Proven Query :
```

```
(ist
  (MtUnionFn UniverseDataMt SportsMt)
  (not
    (behaviorCapable PlanetEarth Marathon doneBy)))
```

```
Rule Assertion :
```

```
●M(implies
  (and
    (isa ?INS ?TYPE)
    (typeBehaviorIncapable ?TYPE ?SITTYPE ?ROLE))
  (behaviorIncapable ?INS ?SITTYPE ?ROLE)) in CapabilitiesMt
```

```
Rule Bindings :
```

```
?TYPE → InanimateObject
?INS → PlanetEarth
?SITTYPE → Marathon
?ROLE → doneBy
```

```
Additional Local Supports :
```

```
:NEGATIONPREDS (negationPreds behaviorIncapable behaviorCapable) in (MtUnionFn UniverseDataMt SportsMt)
```

```
Complete Proof Tree :
```

```
[Proof 6432.6] TRANS-PREDICATE-NEGATIONPREDS-NEG
```

```
●M(implies
  (and
    (isa ?INS ?TYPE)
    (typeBehaviorIncapable ?TYPE ?SITTYPE ?ROLE))
  (behaviorIncapable ?INS ?SITTYPE ?ROLE)) in CapabilitiesMt
:NEGATIONPREDS (negationPreds behaviorIncapable behaviorCapable) in (MtUnionFn UniverseDataMt SportsMt)
```

```
[Proof 6432.5] Join
```

```
[Proof 6432.4] REMOVAL-ALL-ISA
```

```
:ISA (isa PlanetEarth InanimateObject) in (MtUnionFn UniverseDataMt SportsMt)
```

```
[Proof 6432.2] REMOVAL-TVA-UNIFY
```

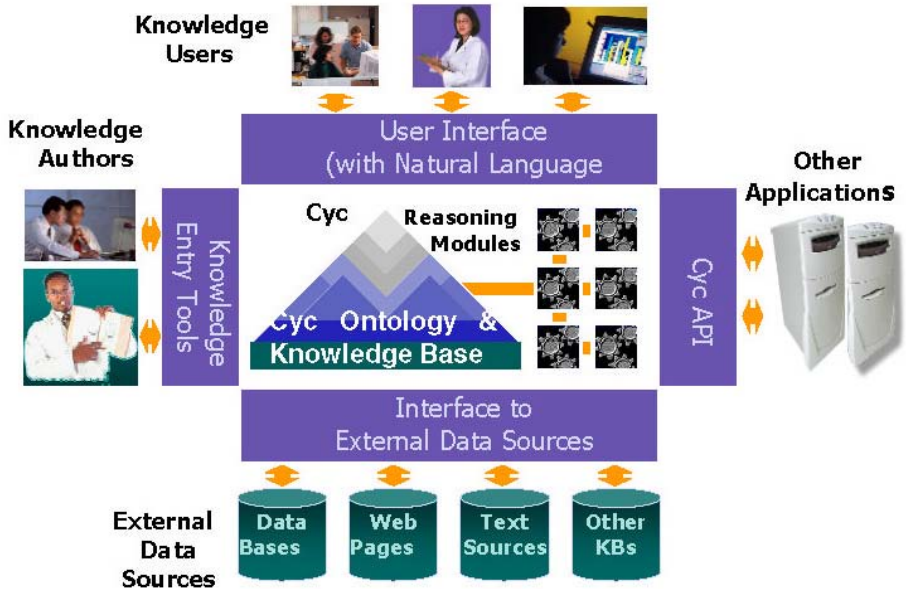
```
:TVA (typeBehaviorIncapable InanimateObject Marathon doneBy) in (MtUnionFn UniverseDataMt SportsMt)
```

**Fig. 2.** Partial Inference Tree for a proof that the planet Earth is not capable of running a marathon, supported by a proof that, more generally, inanimate objects can't run marathons

the world). Knowledge must be hand-crafted and entered into Cyc; it is knowledge that allows Cyc to even begin to understand data. Over the past few years, Cycorp has developed components, jointly referred to as SKSI (Semantic Knowledge Source Integration), which allow knowledge and data to each be stored and accessed in an appropriate way (Masters and Gungördü 2003) [11].

Just as a human researcher would not bother to memorize certain facts, many kinds of data are best kept out of the KB proper. Storage constraints are one reason: although memory is progressively less expensive and more available, a true common-sense system will require an amount of information that is vast even by modern standards. Why use up precious memory to reify every brand-name product sold in the world? As well, many types of data are unstable or ephemeral, varying significantly over time. One simple example is stock prices. Telling Cyc about the stock price of each company on the New York Stock Exchange would be essentially pointless, since that information would rapidly become stale. Rather than continually updating that data within the KB, it would be much more appropriate to consult that data at the source, in real time, so Cyc can always use the most up-to-date figures. These concerns were the motivation for SKSI, which allows Cyc to access data in structured sources, such as databases, and reason with the facts it finds in exactly the same way it reasons with its own native knowledge.

Cyc’s SKSI technology allows the meanings of database columns (for example, “This column represents a person’s office phone number”) to be described in CycL. Only the *interpretation* of database information is represented within Cyc; all the actual data remains in its native format, in the original database. Cyc, in effect, knows how to generate SQL queries to access any particular field in any particular row, and how to interpret what it finds there. These mappings allow Cyc to combine information from multiple databases, or from a combination of databases plus the Cyc KB itself, to answer user queries. Many current Cyc applications assume the need to consult a variety of sources, such as the KB itself; structured data in databases; and unstructured data, such as news articles and Web pages.



**Fig. 3.** Overview of Cyc. The complete Cyc system has users who retrieve knowledge in a variety of ways, and authors who generate knowledge (and who overlap with users). The possible interfaces to the Cyc KB and reasoning capabilities – both those based on NL and more programmatic interfaces to other applications and data sources – are shown on the right and bottom.

## 2 The Case for an Ambient Research Assistant

Performing scientific research is a knowledge-intensive task. Locating, maintaining, and organizing the knowledge required for effective investigation, without interfering with the creative process, catalyzes successful research. Lacking truly intelligent support, researchers must spend a significant portion of their professional time performing infrastructure tasks, all of which are crucial to the overall process. Researchers must manage their personal information flow, locate related work and sources of related expertise, stay abreast of work going on in the field they are

working in, ensure that critical laboratory and computing infrastructure is in place, perform task tracking, write proposals, write and submit publications, and a score of other functions only tangentially related to the actual cycle of hypothesizing, testing, and revising scientific knowledge. In a broader sense, this is true in every field of human endeavor. Providing better tools and interfaces can reduce the time and effort involved in any given task. Project management software simplifies task tracking, and good word processors make writing papers easier. However, some duties do not lend themselves to the creation of specialized tools – arranging for laboratory infrastructure is a time-consuming role that varies enormously across workplaces. Furthermore, such tools do not eliminate the need to perform the tasks, or the attention absorbed by task switching.

Historically, the only agents capable of reducing that load are fairly highly trained human assistants. The range of tasks such an assistant may be called upon to perform is broad, whereas the capabilities built into even the most useful tools are *deep* – that is, focused on a particular task; while Google Scholar, Cora (McCallum et al. 1999) [14], and CiteSeer (Bollacker et al. 1998) [1] reduce the time spent researching related work, they cannot handle arbitrary queries that draw on real-world knowledge, e.g., when was a particular piece of research performed, what is the relationship among research groups in a particular field – concepts that a competent assistant possessed of common sense and a relatively small amount of domain expertise can handle readily. Semi-formally represented knowledge, such as the use of standardized keywords and publication formats, is a traditional and valuable mechanism for helping scientists access information. Extending that to formally represented knowledge (based on a shared ontology) is an effective way of sharing large bodies of knowledge between groups, and reasoning systems that draw on that knowledge are the logical counterpart to tools that perform well on a single, rigidly defined task. A system that rests on a large, non-domain-specific knowledge base can potentially manage tasks that require world knowledge, or common sense – the knowledge that every person can reasonably assume every other person possesses. Until that knowledge is fully represented and integrated, tools will continue to be, at best, *idiots savants*.

The underlying philosophy of the Cyc Project is that software will never reach its full potential until it can react flexibly to a variety of challenges. Systems should not only handle tasks automatically, but also actively anticipate the need to perform them. This requires the development of a variety of technologies (script recognition, integration of external knowledge sources, facilitation of appropriate knowledge entry, hypothesis formation, and so on). These technologies must be integrated into a reasoning system that is possessed of a broad base of pre-existing knowledge, such as the Cyc knowledge base (KB), which provides the system with enough information to understand the context in which each task is being performed.

## 2.1 The Role of Assistance

*Flexibility* is key to creating a truly useful assistant. A good assistant is capable of handling arbitrary problems in arbitrary domains with a minimum of instruction. People are particularly well suited to this because they possess a store of real-world knowledge and skills that allows them to rapidly switch from one type of task to another; even a comparatively lightly trained human can find a submission address

and fax a paper to it, schedule a meeting among several people, and respond to external queries about availability. While different duties may demand different levels of expertise in the scientific domain, almost all require some background knowledge as well. Meanwhile, existing computational “assistants” are deep rather than broad; each focuses on solving a particular problem, and is brittle when faced with anything outside that limited area.

Another key characteristic of a useful assistant is *ease of communication*. If describing or spawning off a task is too expensive, in terms of time or cognition, it becomes impractical to hand those tasks off rather than simply performing them. While a particular formal representation might work best when describing a particular type of task (e.g. how to compute variance for an experiment), the best way of communicating information about a variety of different tasks with minimal cognitive load to the researcher is natural language.

Finally, even if the description and training process is initially complex, an assistant must be capable of *learning*. Generally, tasks do not need to be described each time they must be performed; capable assistants can learn from tasks performed successfully, task failures, and analogous functions they already know how to perform. This idea of learning actually describes a wide variety of functions and capabilities:

1. **Deciding what facts to learn.** An assistant system must reason about what knowledge gaps would be most cost-effective to fill in any given context. If a researcher is considering submitting a paper to an upcoming conference, finding submission dates and contact information is likely to be more useful than organizing older work, and should be a higher priority task.
2. **Learning those facts.** The factual gaps should be filled, from available documentation, online sources, and/or communication with the scientist being assisted. In the aforementioned example, the system should set out to learn any missing facts by appropriately querying all its available sources, both online ones and people, starting with the conference web site or call-for-papers and progressing to information that requires some knowledge of the research in question. The submission information may depend on the track to which the paper is being submitted, which requires knowledge of the research topic.
3. **Learning of rules.** Once knowledge is acquired, it is possible to hypothesize general rules. If several conferences have been identified, an assistant might correlate information about each of them and conclude that conferences in some broad field (e.g. machine learning) are often of interest, or that knowing submission dates is often useful. Such a rule can then guide the selection and prioritization of tasks.
4. **Generalizing rules.** Carrying this example through, an effective assistant might learn from one or more identified rules that, for some particular user or researcher, learning and then tracking dates by which some particular action must occur is valuable.
5. **Testing and revision.** The rules, especially the generalized rules, will need to be tested independently of how they were produced. For example, when a general rule about tracking dates is hypothesized, a system might discover after experimentation that it is less helpful to track and remind a user of recurring dates, such as a weekly report that must be made to an overview body. This discovery would force revision (tightening) of the generalized rule.

## 2.2 The Limitations of Human Assistance

Ultimately, the goal of a personal assistant is to reduce the time and cognitive effort spent on infrastructure tasks. In some ways, computational systems have the potential to assist researchers at a level that a human assistant could never match. Some tasks are pervasive – it makes little sense to have a human assistant file each piece of email after reading, as the time spent splitting off many tiny tasks is greater than the effort of simply performing each task. An assistant's *availability* at the moment some duty must be performed is crucial. An ambient system, in which a user works within an environment in which some agent is actively trying to classify the behavior the user is engaging in, and perform subtasks, has the potential to assist with the many small tasks that create a burden of day-to-day effort, thereby providing assistance that a human assistant could not.

Another crucial behavior for a non-intrusive assistant – that is, one with minimal cognitive load for the user – is *anticipation* of the needs of the researcher. Ambient software assistants have the potential to classify the behavior the user is engaging in, predict useful subtasks and expected future tasks, and either perform those tasks or perform introductory steps before they are required, thus obviating the need for the researcher to identify and describe tasks. This requires *plan recognition*, identifying a user's actions as part of a script, which is either predefined (as in “reading email, then responding, then filing”) or generated on the fly by the system. A script describing reading a research paper may include following reference links and seeking deeper definitions of key terms. If this script is recognized, the system might display recognized terms and links from those terms to other relevant ontologized knowledge while the user is still reading; a user might then be presented with a knowledge acquisition interface to define unrecognized terms, expanding the knowledge base.

## 2.3 Components of a Truly Intelligent Computational Assistant

**Plan recognition:** Creating a truly intelligent assistant will require substantial computational infrastructure. A crucial piece will be a component responsible for gathering information about how people actually perform certain tasks, in as much detail as possible; this is a prerequisite for figuring out how to automate pieces of those tasks. Similarly important is the capacity to *recognize* what a person is trying to do, and to *generate* new scripts to help a user optimize his workflow. All of these actions must occur more or less transparently to the user of the system; otherwise the cognitive load introduced by the intrusiveness of the tool will render it unusable.

**Learning:** Making truly intelligent use of any information collected, and minimizing the effort involved in using the system, requires many different kinds of *learning*, ranging from learning facts and rules to identifying patterns that can serve as a basis for script recognition. Although human training and reinforcement can be involved to some extent, especially with respect to reviewing the system's conclusions, the majority of this learning must perforce take place automatically. Taking advantage of the inferential capabilities present in Cyc allows the automatic or semi-automatic



conjecture of facts, collection of new facts, and production of hypothetical rules and scripts that can then be generalized, tightened, corrected, and used.

**Natural Language:** Finally, natural language understanding and generation are required for optimal interaction – a true assistant would be capable of handling aspects of full discourse processing. An assistant system must be able to remember questions, statements, etc. from the user, and what its own response was, in order to understand the kinds of language ‘shortcuts’ people normally use in context. Input from users will not always be in the form of full sentences or questions. The assistant will need to use the context of what has been said before, along with knowledge of the user’s goals, to interpret requests or commands. For example, a researcher might ask “Can you find me any articles by M. Smith of Stanford on bioethics?” The assistant might then return a list of 200 such articles. The user might reasonably then ask, “Just show me the ones from 2001 or later”, or “Which ones was he the main author on?” In order to seamlessly handle this interaction, the system needs to be able to interpret references (like “the ones” or “he”) to objects already present in the discourse space.

Background on the history, goals, and current state of the Cyc Project were given in Section 1; what follows is an overview of work being done at Cycorp on each of these three high-level components: natural language processing, learning, and ambient interaction.

### 3 Natural Language Processing in Cyc

Spoon-feeding knowledge into Cyc (creating terms and hand-ontologizing each fact about each concept) is time-consuming, and Cyc’s ontological engineers must become adept at translating back and forth between English and CycL. This is a tedious mechanism for knowledge acquisition, and furthermore does not allow free interaction with users who are not trained in CycL. A few years into the Cyc Project, work began on limited natural language processing using Cyc as a substrate, though the plan was always to focus more on NLP during Cyc’s second decade of development.

#### 3.1 Components of Natural Language in Cyc

Natural language understanding (NLU) and natural language generation (NLG) capabilities allow users to interact with Cyc using English instead of CycL. With these capabilities, Cyc can start down the road toward being able to read texts and learn new information on its own.

##### 3.1.1 Cyc’s Lexicon

At the core of Cyc’s natural language processing capabilities is its English lexicon (Burns and Davis 1999) [4]. This lexicon contains information about the syntactic properties of words and phrases (e.g. “tree” is a noun; “eat” has both transitive and intransitive uses), as well as a compositional morphological representation for derived words (for example, “flawless” is decomposed into the root “flaw” plus the suffix “less”). Most important, though, are the semantic links: pointers from words and

phrases to concepts and formulas in the Cyc KB. It is these links that allow for translation of English sentences into fully formed CycL representations, and vice versa. Cyc’s lexicon currently contains entries for over 20,000 single-word noun, verb, adjective, and adverb forms; 40,000 multi-word phrases; and more than 100,000 proper names. Following are partial lexical entries for the words “tree” and “eat”:

***Lexical Information for “tree”***

- CycL: (#\$denotation #Tree-TheWord #CountNoun 1 #Tree-ThePlant)  
Meaning: #Tree-TheWord is a count noun denoting #Tree-ThePlant
- CycL: (#\$singular #Tree-TheWord "tree")  
Meaning: One singular form of #Tree-TheWord is the string “tree”.

***Lexical Information for “eat”***

- CycL: (denotation Eat-TheWord Verb 1 EatingEvent)  
Meaning: #Eat-TheWord is a verb which denotes an #EatingEvent
- Semantic translations of ‘eat’:  
(verbSemTrans Eat-TheWord 0 TransitiveNPFrame  
    (and  
        (isa :ACTION EatingEvent)  
        (performedBy :ACTION :SUBJECT)  
        (consumedObject :ACTION :OBJECT)))  
  
(verbSemTrans Eat-TheWord 0 IntransitiveVerbFrame  
    (and  
        (isa :ACTION EatingEvent)  
        (performedBy :ACTION :SUBJECT)))

Cyc’s lexicon makes use of a specialized microtheory structure to represent languages, including various dialects of English as well as other non-English languages (German, French, Spanish, etc.). Small numbers of words and phrases in these other languages have been added, mainly as a proof-of-concept that Cyc’s lexical representation vocabulary can handle or be easily extended to handle a variety of languages. The current focus is on parsing and generating English, though projects involving other languages, such as Chinese, are underway (Schneider et al. 2005) [17].

### **3.1.2 Natural Language Generation**

Cyc’s NLG system produces a word-, phrase-, or sentence-level paraphrase of KB concepts, rules, and queries. This system relies on information contained in the lexicon, and is driven by generation templates stored in the knowledge base. These templates are not solely string-based; they contain linguistic features that allow, for example, a variety of verb tenses, and correct grammatical agreement, to be produced. The NLG system is capable of providing two levels of paraphrase, depending on the demands of the application. One type of generated text is terse but potentially ambiguous, while the other is more precise, but potentially wordy and stilted. Automated interface tools assist users in adding new generation templates as they introduce new concepts into the knowledge base.

### *Generation for the predicate # $\$$ hasDiet*

- CycL Template:
 

```
(genTemplate hasDiet
  (PhraseFormFn NLSentence
    (ConcatenatePhrasesFn
      (BestDetNbarFn-Indefinite (TermParaphraseFn :ARG1))
      (BestNLWordFormOfLexemeFn-Constrained Adverb
TypicalTheWord)
      (BestHeadVerbForInitialSubjectFn Eat-TheWord)
      (BestDetNbarFn-Indefinite (TermParaphraseFn :ARG2))))))
```
- CycL: (# $\$$ hasDiet # $\$$ Termite # $\$$ Wood)
- Generated Text: “Termites typically eat wood.”

### 3.1.3 Natural Language Understanding

With regard to NLU, depth of parsing from natural language to CycL can range from very shallow (for example, simply mapping strings to concepts) to deep (full text understanding, via translation to CycL formulas). Cyc-based applications have differing needs with respect to parsing speed, depth, and accuracy. This has resulted in the development of a number of in-house parsing tools, including standard CFG parsers and template parsers. External parsing tools, such as Charniak’s (2001) statistical parser and the LINK parser developed at Carnegie-Mellon University (Sleator and Temperley 1991) [20], have also been adapted and used with Cyc’s semantic translation tools.

#### *Parsing Example*

- English: “Bill Clinton sleeps.”
- Parsed CycL:
 

```
(#$thereExists ?SLEEPS
  ($and
    ($isa ?SLEEPS # $\$$ Sleeping)
    ($bodilyDoer ?SLEEPS # $\$$ BillClinton)))
```

### 3.2 Question-Answering Via Natural Language

Cycorp has developed a Cyc-based natural-language question-answering application, designed to operate over heterogeneous text and structured data. The goal of this work is to enable the proactive seeking out of sets of facts that may be relevant to any given task that a user is pursuing, such that that information can be presented in a clear, coherent context. Ultimately, it will assist users in the task of reasoning systematically about entities of interest, hypotheses and facts about those entities, and the likely explanations of and consequences of those hypotheses and facts. The resulting compilations are not just static shoeboxes for facts; they support a computing environment that dynamically recommends actions for researchers to consider. A compilation of facts about a particular conference, for example, might include all relevant submission dates, location, organizers, conference topic, and second-order information such as hotels found in that location. This allows Cyc to suggest

appropriate actions at each stage, and even to initiate actions on the user's behalf. In the long term, the goal of this work is to build infrastructure that assists the general capability of the system to manage an information-gathering task, by providing the means for a user to ask specific questions via natural language. Below is an idealized use case pertaining to the example of scientific conferences:

1. *Upon encountering the expression “the proceedings of the ILP 2004 conference” in a scholarly context, the QA system tries to determine that ILP is a conference series, held in (at least) the year 2004, which publishes formal proceedings.*
2. *Based on this information, the information-gathering system suggests that it might be worth finding out where the conference was held and what the main theme of the conference was. This in turn may lead to questions and/or inferences about who participated and what papers were presented. All such queries are optional – the user can ignore them all if he or she wishes.*
3. *If the user does choose to pursue one of these, the system may suggest several alternate paths, each consisting of several subtasks. For example, suppose the researcher selects a query about whether it is likely that inductive logic programming is the main topic of the conference. One way to handle this query is to search recent articles linking inductive logic programming to known researchers in this field, infer who might have presented at such a conference, and search citation sites for evidence that one or more such persons did indeed present at the conference in question.*
4. *The subtasks then decompose into finding researchers in a particular field, searching citation records for papers involving either a set of those individuals or one of them with a very unusual name, etc.*

Making such suggestions would be of enormous value, and stretches the limits of current knowledge-based systems technology. The next step – which stretches the limits of current natural language parsing and understanding technology – is for the NL system to automatically read through the online documents and extract these subtask answers, after which they can be combined logically to produce an answer. An important feature of the information-gathering system is its *Fact Sheet facility*, which provides a framework for organizing and managing information about entities and events of interest. The formal representations used in the Fact Sheets allow for easy sharing of information across users, and for automated queries to be run against the information in the Fact Sheets. Some of the information in Fact Sheets can be gathered and verified automatically, before being presented to a user (Schneider et al. 2005) [17]. Included in Fact Sheets is meta-data about the provenance (both which document a fact came from, and who interpreted the document) of pieces of information in the Fact Sheet (see Figure 4).

As implemented, the initial step is to gather facts about an entity. This could be an entity that a researcher is specifically interested in, or an entity which Cyc has determined, based on what it knows about the user's interests, would likely be relevant to the task. The system first determines that entity's type. If the entity is already known to Cyc, type data will be extracted from the KB; if not, the system searches using the entity's name (currently it searches using Google), gathering sentences from the returned documents that mention the entity. To obtain a coarse typing – sorting into the kinds of categories one would expect from a traditional Information Extraction system such as FASTUS (Hobbs et al. 1997) [7] or ALEMBIC (Day et al. 1997) [6] – the sentences are run through third-party named-entity recognizers (Klein et al. 2003) [8], (Prager et al. 2000) [15]. Once a rough typing is

obtained (e.g. the entity has been determined to be a person, or a place), syntactic patterns in the retrieved sentences are analyzed in order to refine that typing. This identifies, for example, appositive phrases such as “\_\_\_, a German pharmaceuticals researcher”; these phrases are then interpreted semantically using Cyc’s lexicon.

| Description                   | Fact                                  |
|-------------------------------|---------------------------------------|
| <b>Title:</b>                 | Corpus-Based Knowledge Representation |
| <b>Format:</b>                | conference article                    |
| <b>Appears in book:</b>       |                                       |
| <b>Appears in periodical:</b> | IJCAI 2003 proceeding                 |
| <b>Publisher:</b>             |                                       |
| <b>Date of publication:</b>   | August, 2003                          |
| <b>Topic:</b>                 | Knowledge Representation              |
| <b>Topic:</b>                 | artificial intelligence               |
| <b>Field:</b>                 | artificial intelligence               |
| <b>Author:</b>                | Alon Halevy                           |
| <b>Author:</b>                | Jayant Madhavan                       |
| <b>Editor:</b>                |                                       |
| <b>State of paper:</b>        |                                       |

Publication

**Fig. 4.** Information about a specific paper. Relevant KB content is generated into English and displayed in an editable format. The “green light” metaphor for each assertion shows that there are no consistency issues with the information displayed.

Once an entity has been typed, the next step is to determine which kinds of facts should be gathered for that entity. There are three reasons why a fact might be relevant for research: (1) a user requests it; (2) the system is aware that it is an appropriate type of information (e.g. because an ontologist asserted that it is an appropriate type of information for a particular type of entity, or because knowing that fact will trigger interesting inferences); and (3) it is a type of information that is commonly known for that type of entity. When performing automatically-guided fact gathering about a particular entity, the Cyc system finds out which facts are relevant by consulting existing Fact Sheets, which were created (some manually and some automatically) on the basis of reasons (2) and (3). Next, the system constructs search strings suitable for use by an information retrieval engine. If the search engine finds results, the portions of the resulting sentences that correspond to the blanks in the search strings are semantically analyzed, and the results substituted into the variable position(s) in the original CycL query. Before actually asserting the resulting formulas into the knowledge base, Cyc attempts to verify the proposed facts, using KB consistency checks and additional corpus checks. Finally, verified facts are added to the Fact Sheet, along with meta-information about their sources. (This approach to knowledge acquisition is described more fully in the discussion of learning in Cyc in section 4.2.2.)

After preliminary testing, we believe that this system shows substantial promise as a tool that researchers can use to gather and manage information. Because a formal representation underlies the data stored in the system, others can readily reuse its knowledge (unlike text documents), and automatic updates made by one user can be automatically disseminated to other users. We expect that both the fact-gathering ability and the verification methods will improve as we extend and refine our initial solutions to these problems.

Depending on the particular domain of interest, the fact-gathering system produces results that have a precision level between 50% and 70% (Matuszek et al.; Schneider et al. 2005) [12,17]. The level of fact acquisition is lower; these same tests show that the system finds the sought-after information anywhere from 8% to about 20% of the time. This relatively low number is not surprising, given that the system is asking about entities that in many cases are not well known (e.g., it would not be surprising for no web page to list the organizing committee for a departmental colloquium series). Additionally, the techniques currently being employed are very shallow (requiring exact string matches); the addition of more sophisticated NLP methods should allow for substantially better retrieval rates.

## 4 Learning Within Cyc

In the early days of the web, the “Ask Jeeves” site offered a very exciting prospect: millions of people would pose questions, and hundreds of librarians would find the answers and add them to the site. Over time, the system would become increasingly comprehensive, gradually accumulating the knowledge it would need to answer almost any question.

Nearly a decade after the founding of “Ask Jeeves” in 1996, it has become clear that one of the biggest hurdles facing projects that try to store information with only minimal understanding of the background concept is one of *combinatorics*. The following question is typical, reasonable, and should be answerable; but it will probably never be repeated this century: “What time tracking programs for a Palm Pilot can track on quarter hour intervals, track at least 10 projects, and synchronize with Excel?” Tens of thousands of librarians could not hope to anticipate even a fraction of questions that users asked. The situation is even worse in scientific applications where almost *all* questions are likely to be unique.

Given the availability of a large knowledge base, and the ability to augment that knowledge base using information obtained from the web, it should be possible to successfully direct the learning of new knowledge that in turn improves the system’s ability to anticipate and answer the needs of scientists. We have hypothesized two ways for Cyc to use what it already knows to “bias” or guide this learning: (1) guiding the gathering of facts needed to answer queries (or sub-queries), automatically or via directed dialogues with knowledge workers; and (2) guiding the induction of new knowledge from what is already known.

Performing machine learning over the knowledge centralized in the Cyc knowledge base (or accessible from that knowledge base, as in the integration of relational databases) requires the application of a variety of techniques that are normally used separately. Cyc’s inferential capabilities allow the automatic or semi-

automatic conjecture of facts, the collection of new facts, and the production of hypothetical rules and scripts that can then be generalized, tightened, corrected, and used. This iterative process starts with the low-hanging fruit of implicational rules, and can be expanded to acquisition of more and more complicated knowledge structures. Cyc is a good testing ground for this use of knowledge collection and induction; its large pre-existing corpus of facts and rules provides models that ease the process of fact acquisition and rule induction.

#### 4.1 Learning in Cyc: Goals

Current work in the Cyc Project is taking steps towards beating the combinatorics problem mentioned earlier in relation to Ask Jeeves, creating something like an Ask Jeeves that can directly answer questions. There are two parts to this proposed solution:

1. Use a large knowledge-based system, relying on Cyc, as a representational interlingua, so that  $n$  fragments of information from  $m$  sources can arithmetically and logically combine into answers for novel questions. To do this, the *meaning* of the  $n$  fragments must be available to Cyc. In the specific case of automated (but imperfect) extraction of desired facts from text corpora,  $n$  can be very large.
2. Apply a combination of statistical, deductive, and inductive techniques, to get Cyc to *learn* to answer questions. Some of this learning will be 100% automatic, such as inducing rules from specific facts and generalizing existing rules. Some of this learning will be semi-automatic, such as automatically identifying a small number of specific “pivotal” questions for human users to answer.

Consider even the relatively simple question of booking travel for a conference. To manage this task, the system needs to be able to go out to the web and determine the location of the conference; it needs to know things about the person who is doing the traveling; it needs to be aware of when the travel must occur, which requires sophisticated temporal understanding; and it should have heuristics (rules of good judgment) that specify that for uncertain dates it should volunteer the argument it used to make its guess, but for true specific dates, it should only show the argument if prodded.

The first stages of the necessary learning can be subdivided into two categories: *fact gathering*, the collection of basic knowledge that translates into simple CycL assertions, and *rule production*, which allows the system to conclude to higher-level knowledge given those facts. Fact gathering currently targets ground-level assertions, either at the instance level or the type level, while rule production has focused on generating rules from large sets of ground facts via induction.

**Instance-Level Fact:** (isa Lenat ArtificialIntelligenceResearcher)

**Type-Level Fact:** (genls ArtificialIntelligenceResearcher  
ComputerScientist)

*An AI researcher is a kind of computer scientist.*

**Rule:** (implies  
(and

```
(isa ?CONF Conference)
(topicOfIndividual ?CONF ArtificialIntelligence)
(presenter ?CONF ?PER))
(isa ?PER ArtificialIntelligenceResearcher))
```

*By default, presenters at AI conferences are AI researchers.*

| Create Assignment                     |                                     | Find Assignment  | Target collection "predicates" as consequents |
|---------------------------------------|-------------------------------------|--|---|
| Description                           |                                     | Fact   |   |
| <b>Project:</b>                       | <input checked="" type="checkbox"/> | machine learning approaches  |   |
| <b>Preferred name:</b>                | <input checked="" type="checkbox"/> | Target collection "predicates" as consequents  |   |
| <b>Description:</b>                   | <input checked="" type="checkbox"/> | Target the "unary predicates" (created from collections) as consequents to the induction |   |
| <b>Assigned activity type:</b>        | <input checked="" type="checkbox"/> | computer programming   |   |
| <b>Addresses Bugzilla bug report:</b> | <input type="checkbox"/>            |  |   |
| <b>Request is from:</b>               | <input checked="" type="checkbox"/> | Michael Witbrock   |   |
| <i>Requested employee:</i>            | <input checked="" type="checkbox"/> | Cynthia Matuszek   |   |
| <i>Requested number of hours:</i>     | <input checked="" type="checkbox"/> | 40   |   |
| <i>Requested work period:</i>         | <input checked="" type="checkbox"/> | April, 2005  |   |
| <b>Request is from:</b>               | <input checked="" type="checkbox"/> | Michael Witbrock   |   |
| <i>Requested employee:</i>            | <input checked="" type="checkbox"/> | Robert Kahlert   |   |
| <i>Requested number of hours:</i>     | <input checked="" type="checkbox"/> | 2  |   |
| <i>Requested work period:</i>         | <input checked="" type="checkbox"/> | January, 2005  |   |
| <b>Alternate employee:</b>            | <input type="checkbox"/>            |  |   |

Requests

**Fig. 5.** Factivore used for task assignments. In this form, the employee Cynthia Matuszek is being assigned to the task of importing and exporting CycL collections to an induction system as unary predicates, to which they are logically equivalent. The different colors of “lights” indicate assertions that are already made (green) or in the process of being made (blue).

## 4.2 Gathering Facts

### 4.2.1 Gathering Facts Via User Interaction

In previous work, Cycorp developed a general knowledge acquisition interface called the Factivore (Witbrock et al. 2005) [23]. This component of the Cyc system uses templates, represented in CycL and stored in the Cyc KB, to describe “forms” which, if filled in, in English, to the system’s satisfaction, will result in appropriately formalized assertions appearing in the KB.

This allows users who are not trained ontologists to enter information that might not be possible to obtain otherwise, such as the creation of task assignments for a research group. It is ultimately necessary for a scientist to describe that information somehow – it cannot be retrieved from a web page or calculated from implication rules (at least until AI technology has become substantially more sophisticated!). The least intrusive behavior an assistant can display is to remind a researcher of the need



to provide such instructions to a team, and to provide a mechanism that ensures that it can be done with minimal effort expended on tool use.

For the Factive's initial deployment, ontologists asserted the underlying templates into the KB. While this somewhat painstaking process was reasonable in that context, given the relatively small number of entity types to be represented, and the large number of those instances, the situation with an ambient research assistant is quite different. Depending on task exigencies, new entities of *any* type may need to be represented in the KB. This longer-term goal was the reason for storing the templates in CycL in the KB. It is now possible for Cyc to autonomously produce templates for any type of entity for which Cyc knows a few instances. Over time, it is intended that the system will learn to produce and display these forms in context, eliciting the information required to handle the current task properly, storing it in the KB for future use as learned information, and providing an ever-growing base of “ground facts” to use for rule induction (and improved Factive form design).

#### 4.2.2 Gathering Facts Via Web Search

Given the existing infrastructure in Cyc for representing types and parsing simple natural language sentences, combining these capabilities for targeting knowledge collection from the World Wide Web (e.g. via Google) is a natural mechanism for gathering knowledge (Matuszek et al. 2005). The advantages of targeting the web, rather than a human user, are compelling – the user, who need not answer questions or struggle with the system when the information cannot be understood, need expend no cognitive effort. Given the sea of web pages that can be accessed using tools like Google (Brin and Page 1998) [2], Cyc can simply skip pages it cannot understand, with a substantial likelihood of finding alternate web pages that provide the answer. The implementation of this approach decomposes into several subtasks:

- **Generating strings for web searches** and web page counts from partial logical sentences and logical terms using Cyc's natural language lexicon:

Given a partially-bound CycL phrase, such as:

(occupation Lenat ?WHAT)

Return one or more search strings:

“Lenat has been a \_\_\_\_\_”

“Doug Lenat has been a \_\_\_\_\_”

“Lenat is a \_\_\_\_\_”

- **Using Google's public API** to have Cyc access web pages in their ranked order.
- **Identifying a match** in the web page, then interpreting that match into a formal representation that is consistent with the constraints imposed by the logical sentence:

Finding the string “Lenat has been a professor of ...” in a web page

Interpreting into the term #Professor

Substituting into the original CycL: (occupation Lenat Professor)

- **Eliminating bad interpretations** or erroneous websites by semantically verifying the claim against the KB. Cyc would never suggest (`occupation Lenat PrimeNumber`), (which might be incorrectly parsed from the sentence “Lenat is a prime example...”), because `PrimeNumber` is a member of a class known to be disjoint with `occupation type`.
- **Verifying the correctness** of the parse interpretation by using Google to search for the natural-language interpretation of the assertion we have now constructed, and performing a second round of search over those strings, which allows us to reject interpretations that are a result of bad parses, or are simply too broad. One of the strings “Lenat is a professor” or “Lenat has been a professor” produces results, whereas “Doug Lenat has been a paramedic” – resulting from a mis-parse of a page about machine translation – does not.
- **Asserting the CycL Sentence** into the KB, once it has passed all automatic verification tests.

Targeting the automatic acquisition of simple sentences minimizes the difficulty inherent in generating and parsing complex natural language constructs. As well, simple facts are more likely to be described in a single sentence on the web that can be found and parsed. In initial trials, the search and verification process produced sentences that were correct, according to human review, approximately 50% of the time. While this is not adequate for the needs of a fully autonomous system, when combined with the validation provided by a user who is treating the information provided as a suggestion, it has the potential to be very useful, c.f. (Witbrock et al. 2005) [23].

### 4.2.3 Using Inference to Generate Facts: Abduction

An efficient approach to generating candidate sentences relies on using the *abductive* reasoning capabilities of the Cyc inference engine. In the literature on AI and logic programming, an abduction is generally understood to be an argument of the form:

$$\{Ga \wedge [\forall x (Fx \rightarrow Gx)]\} \rightarrow_{abduced} Fa$$

Thus abduction is generally performed as deduction-in-reverse<sup>1</sup>: working backwards from rules that have the desired result in the *antecedent* from *consequents* that are known to be true and relevant.<sup>2</sup> If new abductive rules are desired, they can be

<sup>1</sup> Logically, abduction and deduction-in-reverse are distinguishable (Mayer and Pirri 1996); in practice, this form of abduction is productive when applied to a large knowledge base containing many deductive rules and an inference harness designed to take advantage of those rules.

<sup>2</sup> More formally, finding candidate hypotheses in Cyc is done by querying the inference engine about the truth of observation *o* (the seed query), which will be a CycL formula containing unbound variables, and then treating the generated deductive proof attempts as candidate explanations of *o*. If we have a seed of the form  $Ga$ , and a rule of the form  $(\forall x) (Fx \rightarrow Gx)$ ,  $Ga$  is passed to Cyc’s inferential query mechanism. With one transformation step the tactician finds  $(\forall x) (Fx \rightarrow Gx)$  and transforms the problem to  $Fa$ . It then attempts to prove  $Fa$ . Whether or not  $Fa$  is true, the inference process stores the problem  $Fa$ .  $Fa$  is thus generated as a problem in a deductive inference’s search for a proof of  $Ga$ . Such generated problems are also abduced hypotheses for  $Ga$ . Accordingly, the converse forms of rules that use `implies` are used to generate abductive hypotheses.

written with `implies`, and they will thereby be available for both deduction and abduction.

As an example: given the previously introduced rule, a new *seed query* (for which hypothetical bindings are desired), and pre-existing knowledge about researchers and conferences:

- **Seed query:** `(presenter AAAI-05 ?WHO)`
- **Assertion:** `(isa Witbrock ArtificialIntelligenceResearcher)`
- **Assertion:** `(isa AAAI-05 Conference)`
- **Assertion:** `(topicOfIndividual AAAI-05 ArtificialIntelligence)`
- **Rule:** `(implies  
     (and  
         (isa ?CONF Conference)  
         (topicOfIndividual ?CONF  
           ArtificialIntelligence)  
         (presenter ?CONF ?PER))  
     (isa ?PER ArtificialIntelligenceResearcher))`

An abductive reasoning process matches the assertion to the consequent of the rule, and generates the following defeasible hypothesis:

- **Candidate Hypothesis:** `(presenter AAAI-05 Witbrock)`

The candidate sentences suggested are checked (via inference and specialized well-formedness-checking modules) for consistency with the current knowledge in the KB. They are then evaluated for *inferential productivity*, i.e., the degree to which they are likely to be subsumed by the antecedent of some rule. Any highly inferentially productive statement has the potential to disproportionately increase the knowledge that can be deduced. When actually performing abduction of this sort, either web verification systems or a human knowledge worker evaluates each sentence for truth and plausibility (Matuszek et al. 2005, Witbrock et al. 2005) [12,23]. The results of this evaluation will be used to improve automated filtering of conjectured sentences, as well as to determine whether those sentences should be immediately applied to tasks where assistance is desirable.

### 4.3 Rule Induction

The successful acquisition of a large, consistently structured data set that is well connected to the existing Cyc knowledge base is possible, based on abductive hypotheses, web search, and gathering facts from users via the Factivore. The logical next step in learning is the *induction* of rules.<sup>3</sup> Induction of implication rules is a process of arguing that a rule is a possible explanation underlying a set of correlated

---

<sup>3</sup> Consistency of form minimizes the inference-based data transformation necessary to perform successful induction, and data that is thoroughly connected to the rest of the knowledge base allows for the automatic formation of better and more descriptive rules.

facts (e.g., if every person that is known to be a mother is a female person, perhaps it can be concluded that all mothers are female). An example rule induced from data might be:

$$[Pa \wedge Pb \wedge Qa \wedge Qb \wedge Rb \wedge Sa \wedge \neg(Sb)]$$

$$\xrightarrow{\text{induced}} \{\forall x [Px \wedge Qx \wedge \neg(Rx) \rightarrow Sx]\}$$

Typical inputs to induction (Srinivasan et al. 2003) [21,22] – a list of first-order facts constructed using a set of related predicates – can be readily produced by simple queries to Cyc’s inference engine over some set of predicates, meaning that running induction continuously over large sets of knowledge found in the Cyc KB is limited only by the speed of the induction process. Induction, like abduction, is not guaranteed to produce *sound* assertions; some are true (accurate when applied to future data points that may be introduced), while others are true only over the training set and must be revised when other, conflicting data points are introduced.

Induction was tested in the project management/personnel management domain in the Cyc Knowledge Base, using a combination of the FOIL 6 (Quinlan and Cameron-Jones 1993) [16] and ALEPH systems (Srinivasan 2001) [21], over an initial set of 10 predicates, which had a combined extent of 1,680 assertions. The predicates used for induction were:

|                       |                        |
|-----------------------|------------------------|
| primarySupervisor     | projectManagers        |
| projectParticipants   | projectTasks           |
| assignedEffortPercent | requestedEffortPercent |
| participantIn         |                        |

- (assignedEffortPercent TASK AGT) means that the IntelligentAgent AGT has been assigned the task TASK.
- (primarySupervisor AGT1 AGT2) means that AGT2 is the default directingAgent in any work-related action in which AGT1 is a deliberateParticipant.
- (participantIn AGENT GATHERING) means that AGENT is an intentional participant in the SocialGathering, GATHERING.
- (projectManagers PROJECT MANAGER) means that MANAGER manages PROJECT, an instance of Project.
- (projectParticipants PROJECT AGENT) means that the intelligentAgent AGENT participates in (usually, receives tasks from) the project PROJECT.
- (projectTasks PROJECT TASK) means that TASK is a task sanctioned by the Project PROJECT. Typically, this means that TASK is a sub-task of the overarching task of ‘completing PROJECT.’
- (requestedEffortPercent TASK AGENT) means that an authorized person (usually the project manager of some parent task of TASK) requests that AGENT work on TASK.

A human reviewer then evaluated the rules. For the initial test runs, sets of predicates from a variety of different domains were selected, and approximately 150 rules were produced over those sets. Two human reviewers independently evaluated all rules using a tool specific to that task (Witbrock et al. 2005) [23]. On average, 7.5% of the automatically produced rules were considered good enough to assert into the KB immediately; 35% more were found to need only quick editing to be assertible. The most common editing required was the deletion of extraneous clauses in the antecedent. On average, it took reviewers 7 hours to review 150 rules, meaning a production rate of ~10 rules per hour when minor editing was allowed, about three times the rate at which a senior Cycorp ontologist can produce rules by hand. Inter-evaluator agreement is approximately 90%. This suggests that, even with human evaluation of rules, induction over ontologized data has the potential to be a comparatively efficient way to discover correlations in at least some bodies of assertions. Some examples of rules that were produced by this system follow:

```
(implies
  (and
    (projectParticipants ?PROJECT ?AGENT)
    (primarySupervisor ?AGT2 ?AGENT)
    (projectManagers ?PROJECT ?AGT2))
  (primaryProject ?AGT2 ?PROJECT))
```

*If someone is a participant in some particular project, someone else is that person's primary supervisor; and the second person is the manager of the project; then the first project is probably that person's primary project.*

```
(implies
  (and
    (primaryProject ?AGT ?PROJECT)
    (projectTasks ?PROJECT ?TASK)
    (requestedEffortPercent ?TASK ?AGT ?PRC))
  (assignedEffortPercent ?TASK ?AGT ?PRC))
```

*If someone's primary project has a task that must be performed; and some percentage of that person's time is requested for that task; then they will probably be assigned at that percentage to that task.*

While these relationships may seem obvious to a person (such as a human assistant), they represent fairly deep reasoning, of the sort that an ambient assistant system must handle transparently and quickly. Not only must the system perform induction successfully; it is necessary to reason about the nature of the terms found in the data set. For example, if this information is being drawn via NL parsing from some corpus, such as a status report, the concept of “participant” may not be explicitly defined. People may be described as “participants,” “manager [of],” or “reviewer,” and the generalization of those to the category “project participant” must be made by either a human, or by a system with broad world knowledge.

## 5 Developing an Ambient Cyc Application

This section outlines current efforts to deploy Cyc to assist with information-gathering tasks by human researchers, analysts, ontologists, and others working in information-intensive fields. These efforts focus largely on building an application framework that extends the Cyc system. This framework derives its core capabilities from the Cyc knowledge base and inference engine, which enable automated, context-sensitive reasoning over multi-source data. The features most relevant to the construction of an ambient assistant are highlighted in this section: anticipation of a user's information needs, and hypothesis generation and tracking.

### 5.1 Anticipation of Information Needs

Cyc incorporates two classes of models of users and analytic processes (illustrated in Figure 6). The first, object-level, class is based on the nature of the objects of inquiry (e.g. models of structured interpersonal transactions guide analyses of social networks). The body of prior knowledge (such as knowledge of people, objects, places, actions, and real-world event types) in the Cyc knowledge base underpins models of object-driven analytical processes. The second meta-level class of models derives from principles of effective research techniques. For example, one such default principle is: "If there are  $n$  *prima facie* equally good alternatives to pursue at a particular stage of analysis, it is equally important to pursue each of the  $n$  alternatives". The first class of models will be used to anticipate users' information needs and to proactively fulfill them. The second class of models will be used to build increasingly accurate individual and composite models of researchers and their needs.

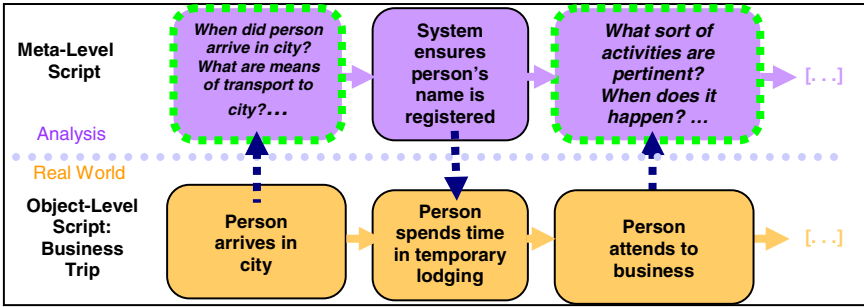
#### 5.1.1 Models of Object-Driven Analysis Processes and Anticipation of Information Needs

Speaking at the level of highest generality, any analytical inquiry about something involves: (1) investigating the thing's parts (using a very broad construal of "parts"), and (2) investigating how the thing bears significant relationships to other things. Cyc understands this feature of analysis by virtue of having a very general script<sup>4</sup> that says "To learn about a thing, one must find out about its significant parts and its significant external relations". (Siegel et al. 2005) [19]

Once realized, an assistant using this analytical schema will know what sorts of fact are relevant when studying certain broad classes of things. For example, when studying an event, such as travel, it will know the importance of understanding what type of event it is in order to situate it in the context of other events – in particular, sub-events and super-events of a focal event like a scholarly conference. Similarly,

---

<sup>4</sup> Scripts are type-level representations of events that have some sort of complexity in terms of either the temporal ordering of their sub-event types (or "scenes"), or in terms of the types of things that typically play roles in their scenes. As such, scripts could enable Cyc to recognize complex actions based on matching sensor data to the patterns of scripts, and they could enable Cyc to perform complex actions according to a default prescribed action sequence.



**Fig. 6.** Analysis Procedure for Business Travel, illustrating the relation between the process of analyzing the behavior of a person and a script that describes the typical progression of a certain class of human behaviors. On the analysis side, finding a person’s name in a hotel register corresponds to a behavior in the real world: a person spending time in a temporary lodging. Spending time in a temporary lodging is typically associated with both travel and a purpose for the travel, each of which are features of the object of analysis, and each of which suggest natural behaviors for an assistant system to pursue.

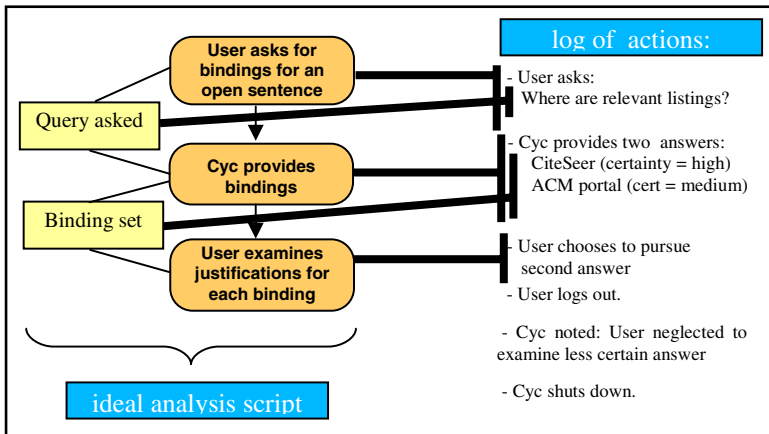
when studying a physical thing, it will try to understand what kinds of things it can be a part of, and what kinds of parts it has. Armed with general principles of this sort, such an assistant will not need to be explicitly told that if an individual is traveling for a conference, it is important that that person be registered in a hotel somewhere in the vicinity of the conference in the time frame in which the conference is taking place. This knowledge will result from Cyc’s understanding of location and co-location, temporality, event occurrence, the meaning of “attendance,” and a host of other factors. Like a good assistant (and unlike much current software), these details can be recognized as related and relevant – not only without the researcher spelling out every detail of his or her needs, but without any explicit interaction between the researcher and the system. By recognizing *classes* of questions as significant, Cyc has motivation to instigate lines of inquiry that will reveal the information (e.g., by accessing transportation schedules), thereby having this information ready for the researcher when the time comes to book travel.

### 5.1.2 Ideal Analysis Processes, Script Learning and the Detection of Bias

The successful realization of Cyc as an ambient assistant relies on having large amounts of data showing what human agents with extensive knowledge requirements actually *do* as they work. The development cycle currently takes advantage of a transaction capturing environment, which allows researchers to capture and record activities occurring during the course of real tasks, including the stream of tasks, queries, documents examined, and reports produced. To the extent possible, it also captures working or draft documents produced. Actions are captured in very fine detail, often down to the keystroke. By interpreting transaction logs and comparing logged behaviors to descriptions of ideal processes, it is anticipated that Cyc will learn new strategies, expose some single-occurrence errors, and detect patterns that indicate important systematic biases on the part of individual users. This knowledge

can then be generalized to broader and broader scripts of human behavior, resulting in knowledge of general scripts that specify how particular tools (such as IR engines, Cyc's own hypothesis generation tools, and other analytical tools) should be used.

For example, a script might be developed over time that says that, when a user is searching the Internet for information about work related to statistical approaches to data retrieval, it is important to execute searches that ultimately incorporate all major relevant research sites – CiteSeer, the ACM portal (<http://portal.acm.org>), etc. The assistant will then be able to detect when a user has deviated from ideal behavior, and suggest a search specifically including a missing site. Discrepancies such as this – unique errors on the part of the researcher – have quick remedies, such as a set of links to IR searches that have already been performed on the remaining relevant terms.



**Fig. 7.** Planned user action tracking. The right-hand side of the diagram displays a list of the actions Cyc and the user take; on the left-hand side is the idealized script. First, the user asks a query and Cyc provides two answers. Next, the user views one of the answers (the more certain one), and then logs out. Cyc logs the fact that the user neglected to pursue one answer, and detects that the behavior logged differs from the learned case.

Ultimately, it will be possible to detect the systematic biases of specific users. By building long logs of the ways behaviors diverge from the ideal in each case, it will be possible to help scientists recognize their own systematic shortcomings, and ultimately make up for those shortcomings. Of course, when there is a discrepancy between Cyc's view of an ideal strategy and a researcher's behavior, it will not always be the user who is at fault. In some cases, a discrepancy will arise because Cyc does not yet know about a particular approach, or simply does not recognize the correct termination case; perhaps the user in the diagram above was not doing a broad literature search, but rather trying to find a specific piece of work, and was satisfied before exhausting the search possibilities.

In some cases, Cyc will potentially be able to *learn* that what may appear to be evidence of bias is actually evidence of effective use of tacit expert knowledge. This would allow the system to distinguish omissions from situations where the user has



yet to perform an action, or already knows what the result of the behavior would be. This kind of sophisticated user model would provide a notable advantage over systems that produce intrusive warnings stating the obvious.

## 5.2 Hypothesis Generation and Tracking

One way in which an assistant can assist a researcher to obtain knowledge is by hypothesizing, in advance, what approaches a researcher would benefit from pursuing. Such hypotheses need to be presented in a context that makes it clear why they should be of interest, and – in order to minimize the cognitive and time cost of tool use – it will be necessary to provide easy one-click tools that enable users to assess hypotheses based on existing documents, and then to confirm or deny the hypotheses. Several such tools have already been developed for a variety of uses within Cyc, ranging from simple sentence reviewers (and more complex rule reviewers) to predicate populators that allow an untrained user to select from a checklist of possible values for an argument position (Witbrock et al. 2005) [23].

Hypotheses are generated, again, by allowing abductive inference over Cyc's common sense and domain-relevant rules. For example, consider an inference initiated during an attempt to find answers for the question: "Who has done work relevant to our current inductive approach to machine learning of rules?" This question can be answered in a fashion analogous to that of the example presented in Section 4.2.3: working backwards from rules that have the desired result in the *antecedent* from *consequents* that are known to be correct and relevant (Siegel et al. 2005) [19].

In addition, work is progressing on a project designed to extend Cyc's capacity to develop hypothetical supports for focused queries into a richer capability called "scenario generation". This scenario generation will be initiated by a description of a "seed event", which users will be able to describe using the Factivore. For example, the system could be tasked with determining in what ways an upcoming conference trip could result in scheduling difficulties. The system will then generate scenarios under which the result could occur – for example, another conference that is likely to be of interest is scheduled for the same block of time, or the person who should attend may be on vacation, or – less probably – the conference might be cancelled. Cyc would generate a range of scenarios ranked in terms of relevance, using rule-clustering technology that ensures reasoning with the most salient rules first. Having generated hypotheses via abduction, Cyc refines the hypotheses, by adding useful information to them via deduction. The next round of abductions will branch again, producing a tree structure among scenario contexts.

## 6 Conclusion

A true AI can fully manage a variety of tasks, not just simplify them or make them faster. Like a good human assistant, a fully realized AI assistant will make tasks silently vanish: you will never be aware that they even needed to happen. Failing that, such an assistant will bring things to your attention only when you must do something about them. Human assistants use common sense in determining which tasks may simply be carried out and then dismissed; on which tasks the supervisor must be kept

apprised of progress; and when to alert the supervisor that some serious roadblock has been encountered. By drawing on a body of knowledge about scientific research, and about such common-sense concepts as what sorts of things motivate agents to act, how time works, and what is or is not a difficult problem, Cyc will be able to carry a part of the cognitive burden of day-to-day scientific research overhead tasks.

Accomplishing this will depend on having many of the same characteristics that are the hallmarks of a human ambient assistant: flexibility, availability, ease of communication, the ability to learn from a variety of sources, and the ability to correlate learned information and learn higher-level information about expectations and priorities. Furthermore, an ambient assistant – one that is part of the environment in which a researcher lives, and can reason about every aspect of the interrelated events and factors that make up the day-to-day life of a researcher – has the potential to do much more, and much less obtrusively, than a human assistant.

Simply developing many different kinds of special-purpose software will not accomplish this extraordinarily challenging vision; it requires common sense, the ability to learn, and deep integration with the tools and tasks a researcher uses daily. While we would hardly claim that Cyc contains all of the information and abilities to achieve an unobtrusive, useful ambient assistant, we have mapped out in this paper a number of abilities that, when combined together, should prove sufficient to provide the basis for such an ambient assistant. Cyc has made substantial strides in each of these areas, and also has an active research program that should enable ever more progress in the future.

## Acknowledgements

Research as wide-ranging and ambitious as the Cyc Project and the many components described in this paper would not be possible without the support of a number of persons and agencies, including AFRL, ARDA, DARPA, NSF, NIST, and others; corporations and research agencies such as Google, which generously allows us to access their API for research such as this; and many other researchers in the field of artificial intelligence who have guided, assisted, and criticized our progress over time.

## References

- [1] Bollacker, K., Lawrence, S., Giles C.L., 1998. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. *Proceedings of the 2nd International Conference on Autonomous Agents*, New York, pp. 116-123.
- [2] Brin, S., Page, L., 1998. Anatomy of a large-scale hypertextual search engine. *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, pp 107-117.
- [3] Buchanan, B.G., Feigenbaum, E.A., 1978. DENDRAL and Meta-DENDRAL: Their applications dimension. *Journal of Artificial Intelligence* 11, 5-24.
- [4] Burns, K., Davis, A., 1990. Building and maintaining a semantically adequate lexicon using Cyc. in: Viegas, E. (Ed.), *Breadth and Depth of Semantic Lexicons*, Kluwer, Dordrecht.
- [5] Copeland, B.J., 2005. Artificial intelligence. *Encyclopædia Britannica Online*, <http://www.britannica.com/eb/article?tocId=9009711>.

- [6] Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson P., Vilain, M., 1997. Mixed initiative development of language processing systems. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., pp. 348-355.
- [7] Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., Tyson, M., 1997. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text, in: Roche, E., Schabes, Y. (Eds.), *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, Massachusetts, pp. 383-406.
- [8] Klein, D, Smarr, J., Nguyen, H., Manning, C., 2003. Named entity recognition with character-level models. *Proceedings of the Seventh Conference on Natural Language Learning*, pp. 180-183.
- [9] Lenat, D. B., Borning, A., McDonald, D., Taylor, C., Weyer, S., 1983. Knoosphere: building expert systems with encyclopedic knowledge. *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 1, pp. 167-169.
- [10] Lenat, D. B., Guha, R.V., 1990. Building Large Knowledge Based Systems. Addison Wesley, Reading, Massachusetts.
- [11] Masters, J., Güngördü, Z., 2003. Structured knowledge source integration: A progress report. Integration of Knowledge Intensive Multiagent Systems, Cambridge, Massachusetts.
- [12] Matuszek, C., Witbrock, M., Kahlert, R.C., Cabral, J., Schneider, D., Shah, P., Lenat, D., 2005. Searching for common sense: Populating Cyc™ from the Web. *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*. Pittsburgh, PA. In Press.
- [13] Mayer, M.C., Pirri, F., 1996. Abduction is not deduction-in-reverse. *Journal of the IGPL*, 4(1): 1-14, 1996.
- [14] McCallum, A., Nigam, K., Rennie, J., Seymore, K., 1999. A machine learning approach to building domain-specific search engines. *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI 1999)*, pp. 662–667.
- [15] Prager, J., Brown, E., Coden, A., Radev, D., 2000. Question answering by predictive annotation. *Proceedings of the 23rd SIGIR Conference*, pp. 184-191.
- [16] Quinlan, R.J., Cameron-Jones, R.M., 1993. FOIL: A midterm report. *Proceedings of the European Conference on Machine Learning*, 667: 3-20.
- [17] Schneider, D., Matuszek, C., Shah, P., Kahlert, R.C., Baxter, D., Cabral, J., Witbrock, M., Lenat, D., 2005. Gathering and managing facts for intelligence analysis. *Proceedings of the 2005 Conference on Intelligence Analysis: Methods and Tools*, McLean, VA.
- [18] Shortliffe, E., 1976. Computer-based Medical Consultations: MYCIN. New York: American Elsevier.
- [19] Siegel, N., Shepard, C.B., Cabral, J., Witbrock, M.J., 2005. Hypothesis generation and evidence assembly for intelligence analysis: Cycorp’s Noöscape application. *Proceedings of the 2005 Conference on Intelligence Analysis: Methods and Tools*, McLean, VA.
- [20] Sleator, D. D., Temperley, D., 1991. Parsing English with a Link Grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University, School of Computer Science, Pittsburgh, PA.
- [21] Srinivasan, A. The Aleph Manual. University of Oxford, [http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph\\_toc.html](http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph_toc.html).
- [22] Srinivasan, A., King, R.D., Bain, M.E., 2003. An empirical study of the use of relevance information in inductive logic programming. *Journal of Machine Learning Research* 4(7): 369-383.
- [23] Witbrock, M., Matuszek, C., Brusseau, A., Kahlert, R.C., Fraser, C.B., Lenat D., 2005. Knowledge begets knowledge: Steps towards assisted knowledge acquisition in Cyc. *Proceedings of the AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors (KVCV)*, Stanford, CA.

# Face for Ambient Interface

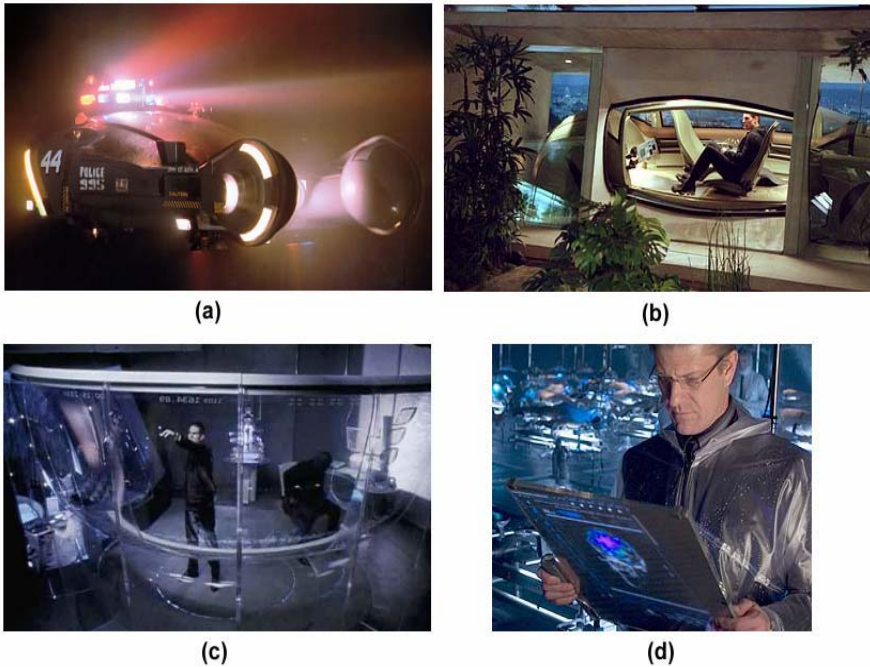
Maja Pantic

Imperial College, Computing Department, 180 Queens Gate,  
London SW7 2AZ, U.K.  
m.pantic@imperial.ac.uk

**Abstract.** The human face is used to identify other people, to regulate the conversation by gazing or nodding, to interpret what has been said by lip reading, and to communicate and understand social signals, including affective states and intentions, on the basis of the shown facial expression. Machine understanding of human facial signals could revolutionize user-adaptive social interfaces, the integral part of ambient intelligence technologies. Nonetheless, development of a face-based ambient interface that detects and interprets human facial signals is rather difficult. This article summarizes our efforts in achieving this goal, enumerates the scientific and engineering issues that arise in meeting this challenge and outlines recommendations for accomplishing this objective.

## 1 Introduction

Films portraying the future often contain visions of human environments of the future. Fitted out with arrays of intelligent, yet invisible devices, homes, transportation means and working spaces of the future can anticipate every need of their inhabitants (Fig. 1). It is this vision of the future that coined the term “ambient intelligence”. According to the Ambient Intelligence (AmI) paradigm, humans will be surrounded by intelligent interfaces that are supported by computing and networking technology embedded in all kinds of objects in the environment and that are sensitive and responsive to the presence of different individuals in a seamless and unobtrusive way [1,40,79]. Thus, AmI involves the convergence of ubiquitous computing, ubiquitous communication, and social user interfaces [64,71] and it assumes a shift in computing – from desktop computers to a multiplicity of smart computing devices diffused into our environment. In turn, it assumes that computing will move to the background, that it will weave itself into the fabric of everyday living spaces and disappear from the foreground [45,74,84], projecting the human user into it [65], and leaving the stage to intuitive social user interfaces. Nonetheless, as computing devices disappear from the scene, become invisible, weaved into our environment, a new set of issues concerning the interaction between ambient intelligence technology and humans is created [74,88]. How can we design the interaction of humans with devices that are invisible? How can we design implicit interaction for sensor-based interfaces? What about users? What does a home dweller, for example, actually want? What are the relevant parameters that can be used by the systems to support us in our activities? If the context is key, how do we arrive at context-aware systems?



**Fig. 1.** Human environments of the future envisioned in motion pictures: (a) speech-based interactive car (*Blade Runner*, 1982), (b) speech- and iris-identification driven car (*Minority Report*, 2002), (c) hand-gesture-based interface (*Minority Report*, 2002), (d) multimedia diagnostic chart and an entirely AmI-based environment (*The Island*, 2005)

One way of tackling these problems is to move from computer-centered designs and toward human-centered designs for human computer interaction (HCI). The former usually involve conventional interface devices like keyboards, mice, and visual displays, and assume that the human will be explicit, unambiguous and fully attentive while controlling information and command flow. This kind of interfacing and categorical computing works well for context-independent tasks like making plane reservations and buying and selling stocks. However, it is utterly inappropriate for interacting with each of the (possibly hundreds) computer systems diffused throughout future AmI environments and aimed at improving the quality of life by *anticipating* the users' needs. The key to ambient interfaces is the ease of use - in this case, the ability to unobtrusively sense the user's behavioral cues and to adapt automatically to the particular user behavioral patterns and the context in which the user acts. Thus, instead of focusing on the computer portion of the HCI context, designs for ambient interfaces should focus on the human portion of the HCI context. They should go beyond the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating social signaling. The design of these functions will require explorations of *what* is communicated (linguistic message, non-linguistic conversational signal, emotion, person identification), *how* the information is communicated (the person's facial expression, head movement, tone of voice, hand

and body gesture), *why*, that is, in which context the information is passed on (where the user is, what his current task is, how he/she feels), and *which* (re)action should be taken to satisfy user needs and requirements.

As a first step towards the design and development of such multimodal context-sensitive ambient interfaces, we investigated facial expressions as a potential modality for achieving a more natural, intuitive, and efficient human interaction with computing technology.

## 1.1 The Human Face

The human face is the site for major sensory inputs and major communicative outputs. It houses the majority of our sensory apparatus: eyes, ears, mouth and nose, allowing the bearer to see, hear, taste and smell. It houses the speech production apparatus and it is used to identify other members of the species, to regulate conversation by gazing or nodding, and to interpret what has been said by lip reading. Moreover, the human face is an accessible “window” into the mechanisms that govern an individual’s emotional and social life. It is our direct and naturally preeminent means of communicating and understanding somebody’s affective state and intentions on the basis of the shown facial expression [38]. Personality, attractiveness, age and gender can also be seen from someone’s face. Thus, the human face is a multi-signal input-output communicative system capable of tremendous flexibility and specificity [24]. In general, it conveys information via four kinds of signals.

1. *Static facial signals* represent relatively permanent features of the face, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals contribute to an individual’s appearance and are usually exploited for person identification.
2. *Slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as the development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual. Note that these signals might diminish the distinctness of the facial features and impede recognition of the rapid facial signals.
3. *Artificial signals* are exogenous features of the face, such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition. Note that these signals might obscure facial features or, conversely, might enhance them.
4. *Rapid facial signals* represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance, including blushing and tears. These (atomic facial) signals underlie *facial expressions*.

All four classes of signals contribute to facial recognition, i.e., person identification. They all contribute to gender recognition, attractiveness assessment, and personality prediction as well. In Aristotle’s time, a theory has been proposed about mutual dependency between static facial signals (physiognomy) and personality: “soft hair reveal a coward, strong chin a stubborn person, and a smile a happy person”<sup>1</sup>. Today,

---

<sup>1</sup> Although this theory is often attributed to Aristotle [4], this is almost certainly not his work (see [4], p. 83).



**Fig. 2.** Type of messages communicated by rapid facial signals. First row: affective states (anger, surprise, disbelief and sadness). Second row: emblems (wink and thumbs up), illustrators and regulators (head tilt, jaw drop, look exchange, smile), manipulators (yawn).

few psychologists share the belief about the meaning of soft hair and strong chin, but many believe that rapid facial signals (facial expressions) communicate emotions [2,24,38] and personality traits [2]. In fact, among the type of messages communicated by rapid facial signals are the following [23,67]:

1. *affective states and moods*, e.g., joy, fear, disbelief, interest, dislike, frustration,
2. *emblems*, i.e., culture-specific communicators like wink,
3. *manipulators*, i.e., self-manipulative actions like lip biting and yawns,
4. *illustrators*, i.e., actions accompanying speech such as eyebrow flashes,
5. *regulators*, i.e., conversational mediators such as the exchange of a look, head nods and smiles.

Given the significant role of the face in our emotional and social lives, it is not surprising that the potential benefits of efforts to automate the analysis of facial signals, in particular rapid facial signals, are varied and numerous, especially when it comes to computer science and technologies brought to bear on these issues. As far as natural interfaces between humans and computers (PCs / robots / machines) are concerned, facial expressions provide a way to communicate basic information about needs and demands to the machine. In fact, automatic analysis of rapid facial signals seems to have a natural place in various vision sub-systems, including automated tools for gaze and focus of attention tracking, lip reading, bimodal speech processing, face / visual speech synthesis, and face-based command issuing. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with

certain commands (e.g., a mouse click) offering an alternative to traditional keyboard and mouse commands. The human capability to “hear” in noisy environments by means of lip reading is the basis for bimodal (audiovisual) speech processing that can lead to the realization of robust speech-driven interfaces. To make a believable “talking head” (avatar) representing a real person, tracking the person’s facial signals and making the avatar mimic those using synthesized speech and facial expressions is compulsory. Combining facial expression spotting with facial expression interpretation in terms of labels like “did not understand”, “disagree”, “inattentive”, and “approves” could be employed as a tool for monitoring human reactions during videoconferences and web-based lectures. Attendees’ facial expressions will inform the speaker (teacher) of the need to adjust the (instructional) presentation.

The focus of the relatively recently-initiated research area of *affective computing* lies on sensing, detecting and interpreting human affective states and devising appropriate means for handling this affective information in order to enhance current HCI designs [61]. The tacit assumption is that in many situations human-machine interaction could be improved by the introduction of machines that can adapt to their users (think about computer-based advisors, virtual information desks, on-board computers and navigation systems, pacemakers, etc.). The information about when the existing processing should be adapted, the importance of such an adaptation, and how the processing/reasoning should be adapted, involves information about how the user feels (e.g. confused, irritated, frustrated, interested). As facial expressions are our direct, naturally preeminent means of communicating emotions, machine analysis of facial expressions forms an indispensable part of affective HCI designs [52].

Automatic assessment of boredom, fatigue, and stress, will be highly valuable in situations where firm attention to a crucial, but perhaps tedious task is essential, such as aircraft and air traffic control, space flight and nuclear plant surveillance, or simply driving a ground vehicle like a truck, train, or car. If these negative affective states could be detected in a timely and unobtrusive manner, appropriate alerts could be provided, preventing many accidents from happening. Automated detectors of fatigue, depression and anxiety could form another step toward personal wellness technologies [20], which scale with the needs of an aging population, as the current medical practices that rely heavily on expensive and overburdened doctors, nurses, and physicians will not be possible any longer. An advantage of machine monitoring is that human observers need not be present to perform privacy-invading monitoring; the automated tool could provide advice, feedback and prompts for better performance based on the sensed user’s facial expressive behavior.

Monitoring and interpretation of facial signals are also important to lawyers, police, security and intelligence agents, who are often interested in issues concerning deception and attitude. Automated facial reaction monitoring could form a valuable tool in these situations, as now only informal interpretations are used. Systems that can recognize friendly faces or, more important, recognize unfriendly or aggressive faces, determine an unwanted intrusion or hooligan behavior, and inform the appropriate authorities, represent another application of facial measurement technology. Systems that adjust music and light levels according to the number, activity, and mood of the users form also an AmI application of this technology.



## 1.2 Why Face for Ambient Interface?

One can easily formulate the answer to this question by considering the breadth of the applied research on AmI and perceptual HCI that uses measures of the face and facial behavior. The preceding section has separately enumerated several computer science research areas and multiple applications in healthcare, industrial, commercial, and professional sectors that would reap substantial benefits from facial measurement technology. This section emphasizes these benefits in the light of the design guidelines defined for ambient interfaces (Table 1).

**Table 1.** The fitness of facial measurement technology for the design of ambient interfaces based upon the design guidelines defined for such interfaces [30,63]

|                                     |  |
|-------------------------------------|--|
| <i>Effective</i>                    | One basic goal for ambient interfaces is continuous provision of background information without disrupting user's foreground tasks. Monitoring the user's attentiveness to the current foreground task based upon his/her facial behavior could help realizing this goal.  |
| <i>Efficient</i>                    | Ambient interfaces should support users in carrying out their tasks efficiently. Examples of how facial measurement technology can help achieving this goal include face-based user identification that relieves users from typing user names and passwords, adapting the amount of the presented information to the level of user's fatigue, and provision of appropriate assistance if confusion can be read from the user's face. |
| <i>easy to learn &amp; remember</i> | It is particularly challenging to achieve ambient interfaces that are easy for the users to learn and to remember, since novel metaphors are used. Nevertheless, the face is the human natural means used to regulate the interaction by gazing, nodding, smiling, frowning, etc. Face-based interfaces would probably be the easiest for the users to "learn and remember".   |
| <i>context-aware</i>                | One basic goal for ambient interfaces is the achievement of the systems' awareness of the context in which the users act [59]: who they are, what their current task is, where they are, how they feel. Face recognition, gaze tracking, and facial affect analysis offer the basis for the design of personalized, affective, task-dependent, natural feeling interfaces.   |
| <i>Control adequate</i>             | It is a particular challenge to realize unambiguous mapping between controls and their effects in the case of ambient interfaces. Note, however, that there should be little (if any) ambiguity about the effect of input facial signals like identity, gaze focus, smiles and frowns, especially when it comes to typically constrained AmI scenarios involving a certain individual's home, car, or work space.                    |
| <i>Domain adequate</i>              | Adequacy of the ambient interfaces for the target domain including the users, their tasks and the environment should be ensured [73]. Although facial measurement technology cannot ensure realization of this goal on its own, it has the potential to accommodate a broad range of users through customized face-based interaction controls for support of different users with different abilities and needs.                     |

---

*participatory design* In the design of ambient interfaces, it is important to stimulate the users to contribute to the design at early stages, so that products tailored to their preferences, needs and habits can be ensured. This is inherent in face-based AmI technology, which relies on machine learning and sees the human user as the main actor. Improving the quality of life by anticipating one's needs via rigid, impersonalized systems is unrealistic; the necessary mapping of sensed facial signals (there are more than 7000 of these [69]) onto a set of controls and preemptive behaviors is by far too complex to be precompiled and hardwired into the system. Systems should learn their expertise by having the user instruct them (explicitly / implicitly) on the desired (context-sensitive) interpretations of sensed facial signals [52].

---

In addition, ambient interfaces should have *good utility* (e.g. they are not suitable for complex information processing) and be *transparent* (i.e. users should be aware, at any time, of what is expected from them, whether the input was received, whether the actions are or will be performed, etc.). Facial measurement technology represents a novel interface modality; it has no direct answers to these basic design questions.

---

While all agree that facial measurement technology has a natural place in AmI technologies, especially in human-centered natural-feeling ambient interfaces, one should be aware of the likelihood that face-based ambient interfaces still lie in the relatively distant future. Although humans detect and analyze faces and facial expressions in a scene with little or no effort, development of an automated system that accomplishes this task is rather difficult. There are several related problems [52]. The first is to find faces in the scene independent of clutter, occlusions, and variations in head pose and lighting conditions. Then, facial features such as facial characteristic points (e.g. the mouth corners) or parameters of a holistic facial model (e.g. parameters of a fitted Active Appearance Model) should be extracted from the regions of the scene that contain faces. The system should perform this accurately, in a fully automatic manner and preferably in real time. Eventually, the extracted facial information should be interpreted in terms of facial signals (identity, gaze direction, winks, blinks, smiles, affective states, moods) in a context-dependent (personalized, task- and application-dependent) manner. For exhaustive surveys of the entire problem domain, the readers are referred to: Samal and Iyengar [68] for an overview of early works, Tian et al. [78] and Pantic [47] for surveys of techniques for detecting facial muscle actions (AUs), and Pantic and Rothkrantz [51,52] for surveys of current efforts. These surveys indicate that although the fields of computer vision and facial information processing witnessed rather significant advances in the past few years, most of the aforementioned problems still represent significant challenges facing the researchers in these and the related fields. This paper summarizes our efforts in solving some of these problems, enumerates the scientific and engineering issues that arise in meeting these challenges and outlines recommendations for accomplishing the new facial measurement technology.

## 2 Face Detection

The first step in facial information processing is face detection, i.e., identification of all regions in the scene that contain a human face. The problem of *finding faces* should be solved regardless of clutter, occlusions, and variations in head pose and lighting conditions. The presence of non-rigid movements due to facial expression and a high degree of variability in facial size, color and texture make this problem even more difficult. Numerous techniques have been developed for face detection in still images [39,87]. However, most of them can detect only upright faces in frontal or near-frontal view. The efforts that had the greatest impact on the community (as measured by, e.g., citations) include the following.

Rowley et al. [66] used a multi-layer neural network to learn the face and non-face patterns from the intensities and spatial relationships of pixels in face and non-face images. Sung and Poggio [75] have proposed a similar method. They used a neural network to find a discriminant function to classify face and non-face patterns using distance measures. Moghaddam and Pentland [44] developed a probabilistic visual learning method based on density estimation in a high-dimensional space using an eigenspace decomposition. The method has been applied to face localization, coding and recognition. Pentland et al. [60] developed a real-time, view-based and modular (by means of incorporating salient features, such as the eyes and the mouth) eigenspace description technique for face recognition in variable pose.

Among all the face detection methods that have been employed by automatic facial expression analyzers, the most significant work is arguably that of Viola and Jones [82]. They developed a real-time face detector consisting of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, which remind of Haar Basis functions and can be computed very fast at any location and scale (Fig. 4(a)). This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost.

There are several adapted versions of the Viola-Jones face detector and the one that we employ in our systems uses GentleBoost instead of AdaBoost. It also refines the originally proposed feature selection by finding the best performing single-feature classifier from a new set of filters generated by shifting and scaling the chosen filter by two pixels in each direction, as well as by finding composite filters made by reflecting each shifted and scaled feature horizontally about the center and superimposing it on the original [27]. Finally the employed version of the face detector uses a smart training procedure in which, after each single feature, the system can decide whether to test another feature or to make a decision. By this, the system retains information about the continuous outputs of each feature detector rather than converting to binary decisions at each stage of the cascade. The employed face detector was trained on 5000 faces and millions of non-face patches from about 8000 images collected from the web by Compaq Research Laboratories [27]. On the test set of 422 images from the Cohn-Kanade facial expression database [37], the most commonly used database of face images in the research on facial expression analysis, the detection rate was 100% [83].

### 3 Facial Feature Extraction

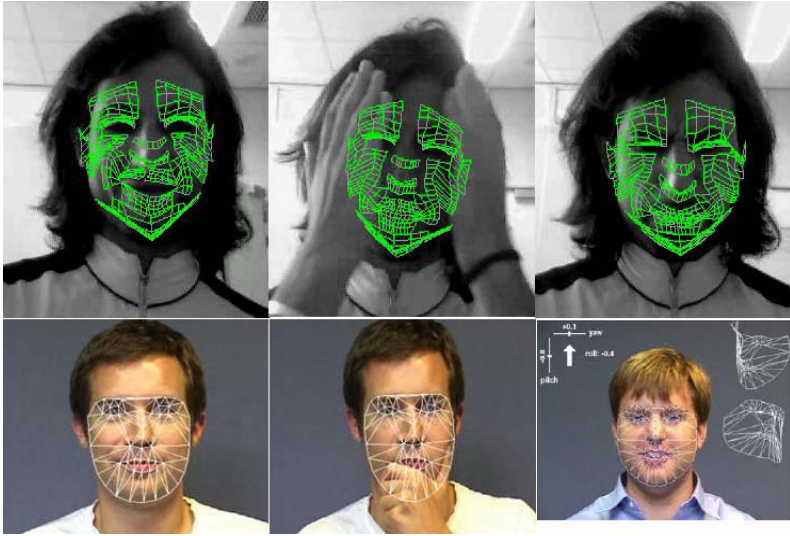
After the presence of a face has been detected in the observed scene, the next step is to extract the information about the displayed facial signals. The problem of *facial feature extraction* from regions in the scene that contain a human face may be divided into at least three dimensions [51]:

1. Is temporal information used?
2. Are the features holistic (spanning the whole face) or analytic (spanning sub-parts of the face)?
3. Are the features view- or volume based (2D/3D)?

Given this glossary and if the goal is face recognition, i.e., identifying people by looking at their faces, most of the proposed approaches adopt 2D holistic static facial features. On the other hand, many approaches to automatic facial expression analysis adopt 2D analytic spatio-temporal facial features [52]. This finding is also consistent with findings from the psychological research suggesting that the brain processes faces holistically rather than locally whilst it processes facial expressions locally [9,13]. What is, however, not entirely clear yet is whether information on facial expression is passed to the identification process to aid recognition of individuals or not. Some experimental data suggest this [42]. Although relevant for the discussion of facial measurement tools and face-based ambient interfaces, these issues are not elaborated further in this paper, as the focus of our past research was mainly automatic facial expression analysis. For exhaustive surveys of efforts aimed at face recognition, the readers are referred to: Zhao et al. [90], Bowyer [12], and Li and Jain [39].

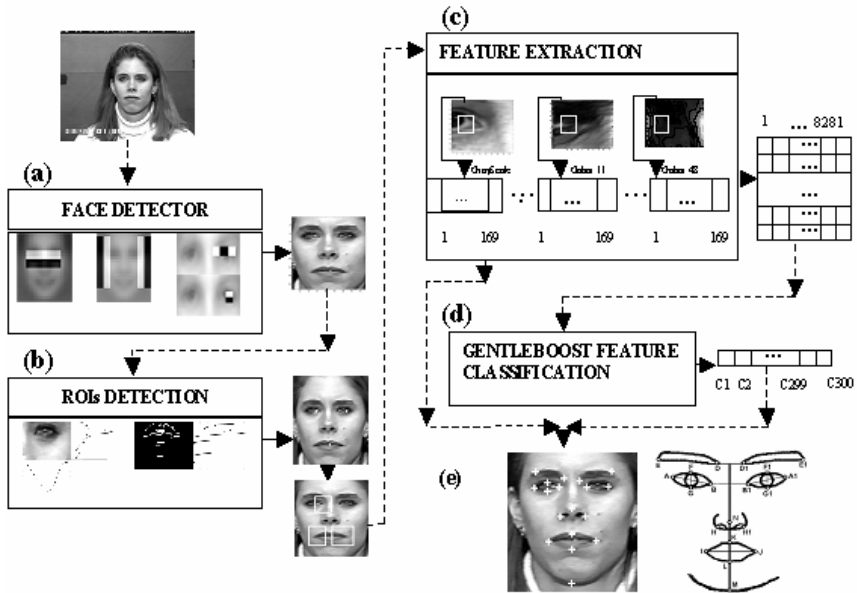
Most of the existing facial expression analyzers are directed toward 2D spatio-temporal facial feature extraction, including the methods proposed by our research team. The usually extracted facial features are either *geometric features*, such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.) or *appearance features* representing the texture of the facial skin, including wrinkles, bulges, and furrows [7,8]. Typical examples of geometric-feature-based methods are those of Gokturk et al. [29], who used 19 point face mesh, and of Pantic et al. [49,53,81], who used a set of facial characteristic points like the ones illustrated in Fig. 10. Typical examples of *hybrid*, geometric- and appearance-feature-based methods are those of Tian et al. [77], who used shape-based models of eyes, eyebrows and mouth and transient features like crows-feet wrinkles and nasolabial furrow, and of Zhang and Ji [89], who used 26 facial points around the eyes, eyebrows, and mouth and the same transient features as Tian et al [77]. Typical examples of appearance-feature-based methods are those of Bartlett et al. [8,21] and Guo and Dyer [32], who used Gabor wavelets, of Anderson and McOwen [3], who used a holistic, monochrome, spatial-ratio face template, and of Valstar et al. [80], who used temporal templates (see Section 3.2).

It has been reported that methods based on geometric features are usually outperformed by those based on appearance features using, e.g., Gabor wavelets or eigen-faces [7]. Recent studies have shown that this claim does not always hold [49,81]. Moreover, it seems that using both geometric and appearance features might be the best choice in the case of certain facial expressions [49].



**Fig. 3.** Examples of 2D and 3D face models. First row: Results of Tao-Huang 3D-wireframe face-model fitting algorithm for happy, occluded and angry face image frames [76]. Second row: Results of the CMU 2D-AAM fitting algorithm for happy and occluded face image frames and results of fitting the CMU 2D+3D AAM [5,85].

Few approaches to automatic facial expression analysis based on 3D face modeling have been proposed recently (Fig. 3). Gokturk et al. [29] proposed a method for recognition of facial signals like brow flashes and smiles based upon 3D deformations of the face tracked on stereo image streams using a 19-point face mesh and standard optical flow techniques. The work of Cohen et al. [14] focuses on the design of Bayesian network classifiers for emotion recognition from face video based on facial features tracked by so-called Piecewise Bezier Volume Deformation tracker [76]. This tracker employs an explicit 3D wireframe model consisting of 16 surface patches embedded in Bezier volumes. Cohn et al. [15] focus on automatic analysis of brow actions and head movements from face video and use a cylindrical head model to estimate the 6 degrees of freedom of head motion. Baker and his colleagues developed several algorithms for fitting 2D and combined 2D+3D Active Appearance Models to images of faces [85], which can be used further for various studies concerning human facial behavior [5]. 3D face modeling is highly relevant to the present goals due to its potential to produce view-independent facial signal recognition systems. The main shortcomings of the current methods concern the need of a large amount of manually annotated training data and an almost always required manual selection of landmark facial points in the first frame of the video-based input on which the face model will be warped to fit the face. Automatic facial feature point detection of the kind proposed in Section 3.1 offers a solution to these problems.



**Fig. 4.** Outline of our Facial Point Detection method [83]. (a) Face detection using Haar-feature-based GentleBoost classifier [27]; (b) ROI extraction, (c) feature extraction based on Gabor filtering, (d) feature selection and classification using GentleBoost classifier, (e) output of the system compared to the face drawing with facial landmark points we aim to detect.

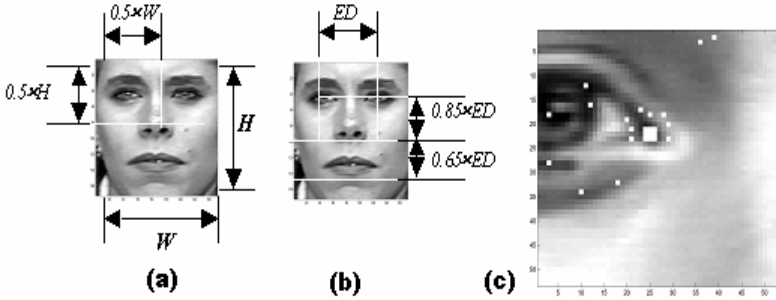
### 3.1 Geometric Feature Extraction – Facial Feature Point Detection

The method that we use for fully automatic detection of 20 facial feature points, illustrated in Fig. 4(e) and Fig 10, uses Gabor-feature-based boosted classifiers [83]. The method adopts the fast and robust face detection algorithm explained in Section 2, which represents an adapted version of the original Viola-Jones detector [27,82].

The detected face region is then divided in 20 regions of interest (ROIs), each one corresponding to one facial point to be detected. The irises and the medial point of the mouth are detected first. The detection is done through a combination of heuristic techniques based on the analysis of the vertical and horizontal histograms of the upper and the lower half of the face-region image achieves this (Fig. 5). Subsequently, we use the detected positions of the irises and the medial point of the mouth to localize 20 ROIs. An example of ROIs extracted from the face region for points B, I, and J, is depicted in Fig. 4(b).

The employed facial feature point detection method uses individual feature patch templates to detect points in the relevant ROI. These feature models are GentleBoost templates built from both gray level intensities and Gabor wavelet features. Recent work has shown that a Gabor approach for local feature extraction outperformed Principal Component Analysis (PCA), the Fisher’s Linear Discriminant (FLD) and the Local Feature Analysis [21]. This finding is also consistent with our experimental data that show the vast majority of features (over 98%) that were selected by the

utilized GentleBoost classifier [28] were from the Gabor filter components rather than from the gray level intensities. The essence of the success of Gabor filters is that they remove most of the variability in image due to variation in lighting and contrast, while being robust against small shifts and deformation [35].

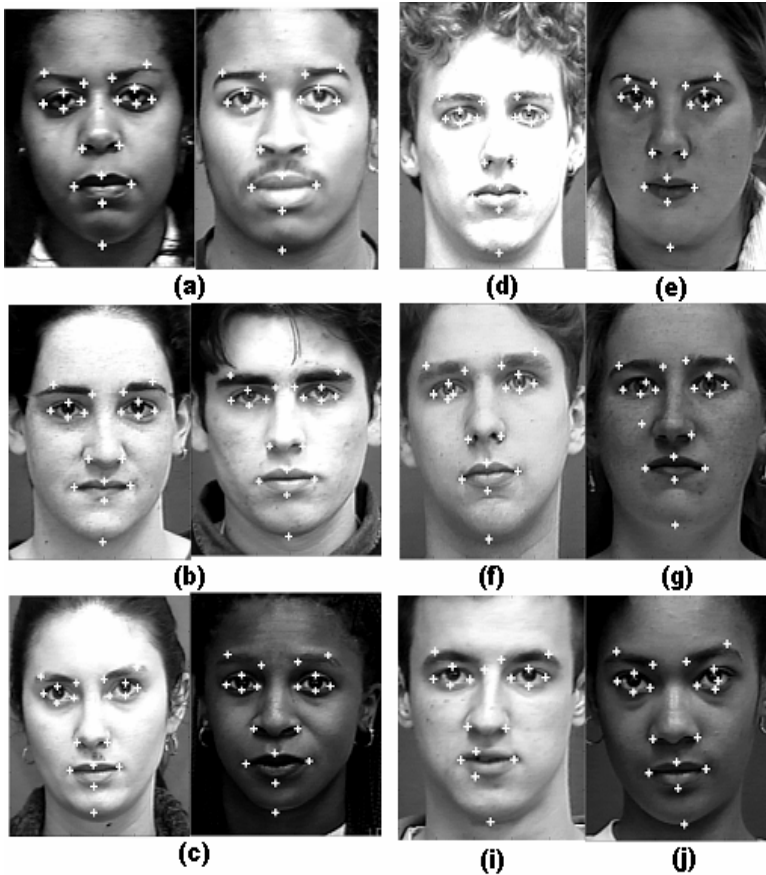


**Fig. 5.** (a) Dividing the face horizontally in half and dividing the upper face region vertically in half. (b) Finding the mouth region within the face region by means of Eye Distance (ED). (c) Positive and negative examples for training point B. The big white square on the inner corner of the eye represents 9 positive examples. Around that square are 8 negative examples randomly chosen near the positive examples. Another 8 negative examples are randomly chosen from the rest of the region.

The feature vector for each facial point is extracted from the  $13 \times 13$  pixels image patch centered on that point. This feature vector is used to learn the pertinent point's patch template and, in the testing stage, to predict whether the current point represents a certain facial point or not. This  $13 \times 13$  pixels image patch is extracted from the gray scale image of the ROI and from 48 representations of the ROI obtained by filtering the ROI with a bank of 48 Gabor filters at 8 orientations and 6 spatial frequencies (2:12 pixels/cycle at  $\frac{1}{2}$  octave steps). Thus,  $169 \times 49 = 8281$  features are used to represent one point. Each feature contains the following information: (i) the position of the pixel inside the  $13 \times 13$  pixels image patch, (ii) whether the pixel originates from a grayscale or from a Gabor filtered representation of the ROI, and (iii) if appropriate, which Gabor filter has been used (Fig. 4(c)).

In the training phase, GentleBoost feature templates are learned using a representative set of positive and negative examples. As positive examples for a facial point, we used 9 image patches centered on the true point and on 8 positions surrounding the true (manually labeled) facial point in a training image. For each facial point we used two sets of negative examples. The first set contains 8 image patches randomly displaced 2-pixels distance from any of the positive examples. The second set contains 8 image patches randomly displaced in the relevant ROI (Fig. 5).

In the testing phase, each ROI is filtered first by the same set of Gabor filters used in the training phase (in total, 48 Gabor filters are used). Then, for a certain facial point, an input  $13 \times 13$  pixels window (*sliding window*) is slid pixel by pixel across 49 representations of the relevant ROI (grayscale plus 48 Gabor filter representations). For each position of the sliding window, the GentleBoost classification method [28]

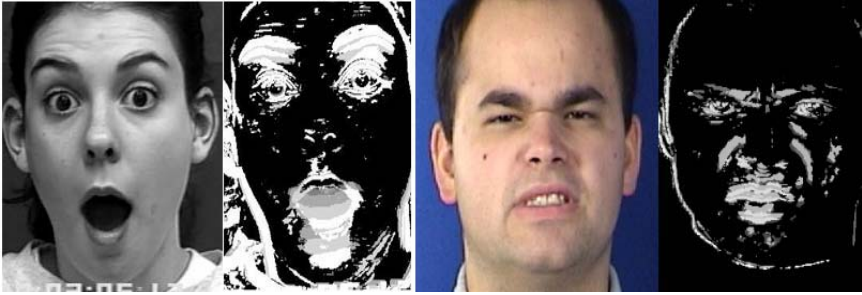


**Fig. 6.** Typical results of our Facial Feature Point Detector [83]. (a)-(c) Examples of accurate detection. (d)-(j) Examples of inaccurate detection of point D1 (d), point E (e), point B (f), points F and H (g), point D and K (i), and point A1 (j). For point notation see Fig. 12.

outputs a response depicting the similarity between the 49-dimensional representation of the sliding window and the learned feature point model. After scanning the entire ROI, the position with the highest response reveals the feature point in question.

We trained and tested the facial feature detection method on the Cohn-Kanade facial expression database [37]. We used only the first frames of the 300 Cohn-Kanade database samples. No further registration of the images was performed. The 300 images of the data set were divided into 3 subsets containing 100 images each. The proposed method has been trained and tested using a leave-one-subset-out cross validation. To evaluate the performance of the method, each of the automatically located facial points was compared to the true (manually annotated) point. As explained above, we used as positive examples the true location of the point and 8 positions surrounding the true facial point in a training image. Hence, automatically detected





**Fig. 7.** Original maximal activation (apex) frames and the related Motion History Images (MHI). Left: AU1+AU2+AU5+AU27 (from the Cohn-Kanade facial expression database). Right: AU9+AU10+AU25 (from the MMI facial expression database).

points displaced 1-pixel distance from relevant true facial points are regarded as SUCCESS. Additionally, we define errors with respect to the inter-ocular distance measured in the test image (80 to 120 pixels in the case of image samples from the Cohn-Kanade database). An automatically detected point displaced in any direction, horizontal or vertical, less than 5% of inter-ocular distance (i.e., 4 to 6 pixels in the case of image samples from the Cohn-Kanade database) from the true facial point is regarded as SUCCESS. This is in contrast to the current related approaches developed elsewhere (e.g. [17]), which are usually regarded as SUCCESS if the bias of automatic labeling result to the manual labeling result is less than 30% of the true (annotated manually) inter-ocular distance.

Overall, we achieved an average recognition rate of 93% for 20 facial feature points using the above described evaluation scheme. Typical results are shown in Fig. 6. Virtually all misclassifications (most often encountered with points F1 and M) can be attributed to the lack of consistent rules for manual annotation of the points. For details about this method, the readers are referred to Vukadinovic and Pantic [83].

### 3.2 Appearance Feature Extraction – Temporal Templates

Temporal templates are 2D images constructed from image sequences, which show motion history, that is, where and when motion in the input image sequence has occurred [11]. More specifically, the value of a pixel in a Motion History Image (MHI) decays over time, so that a high intensity pixel denotes recent motion, a low intensity pixel denotes a motion that occurred earlier in time, and intensity zero denotes no motion at all at that specific location (Fig. 7). A drawback innate to temporal templates proposed originally by Bobick and Davis [11] is the problem of motion self-occlusion due to overwriting. Let us explain this problem by giving an example. Let us denote an upward movement of the eyebrows as action  $A_1$  and a downward movement of the eyebrows back to the neutral position as action  $A_2$ . Both actions produce apparent motion in the facial region above the neutral position of the eyebrows (Fig. 7). If  $A_2$  follows  $A_1$  in time and if the motion history of both actions is recorded within a single Motion History Image (MHI), then the motion history of action  $A_2$  overwrites the motion history of  $A_1$ ; the information about the motion

history of action  $A_i$  is lost. To overcome this problem, we proposed to record the motion history at multiple time intervals and to construct Multilevel Motion History Image (MMHI), instead of recording the motion history once for the entire image sequence and constructing a single MHI [80].

Before we can construct a MMHI from an input video, the face present in the video needs to be registered in two ways. Intra registration removes all rigid head movements within the input video, while the inter registration places the face at a predefined location in the scene. This transformation uses facial points whose spatial position remains the same even if a facial muscle contraction occurs (i.e., points B, B1, and N illustrated in Fig. 10). The inter registration process warps the face onto a predefined “normal” face, eliminating inter-person variation of face shape and facilitating the comparison between the facial expression shown in the input video and template facial expressions. Under the assumption that each input image sequence begins and ends with a neutral facial expression, we downsample the number of frames to a fixed number of  $(n+1)$  frames. In this way, our system becomes robust to the problem of varying duration of facial expressions.

After the registration and time warping of the input image sequence, the MHI is obtained as follows. Let  $I(x, y, t)$  be an image sequence of pixel intensities of  $k$  frames and let  $D(x, y, t)$  be the binary image that results from pixel intensity change detection, that is by thresholding  $|I(x, y, t) - I(x, y, t-1)| > th$ , where  $x$  and  $y$  are the spatial coordinates of picture elements and  $th$  is the minimal intensity difference between two images. In an MHI, say  $H_t$ , the pixel intensity is a function of the temporal history of motion at that point with  $t$  being a frame of the downsampled input video (with  $(n+1)$  frames). Using the known parameter  $n$ ,  $H_t$  is defined as:

$$H_t(x, y, t) = \begin{cases} s * t & D(x, y, t) = 1 \\ H_t(x, y, t-1) & \text{otherwise} \end{cases} \quad (1)$$

where  $s = (255/n)$  is the intensity step between two history levels and where  $H_t(x, y, t) = 0$  for  $t \leq 0$ . The final MHI, say  $H(x, y)$ , is found by iteratively computing equation (1) for  $t = 1 \dots n+1$ .

With an MMHI, we want to encode motion occurring at different time instances on the same location such that it is uniquely decodable later on. To do so, we use a simple bit-wise coding scheme. If motion occurs at time instance  $t$  at position  $(x, y)$ , we add 2 to the power of  $(t-1)$  to the old value of the MMHI:

$$M(x, y, t) = M(x, y, t-1) + D(x, y, t) \cdot 2^{t-1} \quad (2)$$

with  $M(x, y, t) = 0$  for  $t \leq 0$ . Because of the bitwise coding scheme, we are able to separate multiple motions occurring at the same position in the classification stage.

We utilized further a temporal-template-based face image sequence representation for automatic recognition of facial signals such as brow flashes, smiles, frowns, etc. (i.e., for AU detection). Comparison of two classification schemes: (i) a two-stage classifier combining a kNN-based and a rule-based classifier, and (ii) a SNoW classifier, can be found in Valstar et al. [80]. The evaluations studies on two different databases, the Cohn-Kanade [37] and the MMI facial expression database [56], suggest that (M)MHIs are very suitable for detecting various facial signals. Especially

AU1+AU2 (eyebrows raised), AU10+AU25 (raised upper lip), AU12+AU25 (smile with lips parted) and AU27 (mouth stretched vertically) are easily recognized. However, as it is the case with all template-based methods, for each and every facial signal (new class) to be recognized, a separate template should be learned. Given that there are more than 7000 different facial expressions [69], template-based methods including temporal templates, do not represent the best choice for realizing facial measurement tools.

## 4 Facial Feature Tracking

Contractions of facial muscles induce movements of the facial skin and changes in the appearance of facial components such as eyebrows, nose, and mouth. Since motion of the facial skin produces optical flow in the image, a large number of researchers have studied optical flow tracking [39,51]. The optical flow approach to describing face motion has the advantage of not requiring a facial feature extraction stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Because optical flow is the visible result of movement and is expressed in terms of velocity, it can be used to represent facial actions directly. One of the first efforts to utilize optical flow for recognition of facial expressions was the work of Mase [43]. Many other researchers adopted this approach including Black and Yacoob [10], who used the flows within local facial areas of the facial components for expression recognition purposes. For exhaustive surveys of these methods, the reader is referred to Pantic and Rothkrantz [51] and Li and Jain [39].

Standard optical flow techniques [6,41,72] are also most commonly used for tracking facial feature points. DeCarlo and Metaxas [19] presented a model-based tracking algorithm in which face shape model and motion estimation are integrated using optical flow and edge information. Gokturk et al. [29] track the points of their 19-point face mesh on the stereo image streams using the standard Lucas-Kanade optical flow algorithm [41]. To achieve facial feature point tracking Tian et al. [77] and Cohn et al. [15,16] used the standard Lucas-Kanade optical flow algorithm too. To realize fitting of 2D and combined 2D+3D Active Appearance Models to images of faces [85], Xiao et al. use an algorithm based on an "inverse compositional" extension to the Lucas-Kanade algorithm.

To omit the limitations inherent in optical flow techniques, such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, several researchers used sequential state estimation techniques to track facial feature points in image sequences. Both, Zhang and Ji [89] and Gu and Ji [31] used facial point tracking based on a Kalman filtering scheme, which is the traditional tool for solving sequential state problems. The derivation of the Kalman filter is based on a state-space model [36], governed by two assumptions: (i) linearity of the model and (ii) Gaussianity of both the dynamic noise in the process equation and the measurement noise in the measurement equation. Under these assumptions, derivation of the Kalman filter leads to an algorithm that propagates the mean vector and covariance matrix of the state estimation error in an iterative manner and is optimal in the Bayesian setting. To deal with the state estimation in nonlinear dynamical systems, the extended Kalman filter has been proposed, which is derived through linearization of the

state-space model. However, many of the state estimation problems, including human facial expression analysis, are nonlinear and quite often non-Gaussian too. Thus, if the face undergoes a sudden or rapid movement, the prediction of features positions from Kalman filtering will be significantly off. To handle this problem, Zhang and Ji [89] and Gu and Ji [31] used the information about the IR-camera-detected pupil location together with the output of Kalman filtering to predict facial features positions in the next frame of an input face video. To overcome these limitations of the classical Kalman filter and its extended form in general, particle filters have been proposed. An extended overview of the various facets of particle filters can be found in [33].

The tracking scheme that we utilize to track facial feature points in an input face image sequence is based on particle filtering. The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability  $p(\alpha|Y)$ , where  $\alpha$  is the state of a temporal event to be tracked given a set of noisy observations  $Y = \{y^1, \dots, y^t, y^{\bar{t}}\}$  up to the current time instant. This means that the distribution  $p(\alpha|Y)$  is represented by a set of pairs  $\{(s_k, \pi_k)\}$  such that if  $s_k$  is chosen with probability equal to  $\pi_k$ , then it is as if  $s_k$  was drawn from  $p(\alpha|Y)$ . By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering), particle filtering is able to track multimodal conditional probabilities  $p(\alpha|Y)$ , and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear, non-Gaussian systems. In the particle filtering framework, our knowledge about the *a posteriori* probability  $p(\alpha|Y)$  is updated in a recursive way. Suppose that at a previous time instance we have a particle-based representation of the density  $p(\alpha^-|Y^-)$ , i.e., we have a collection of  $K$  particles and their corresponding weights (i.e.  $\{(s_k^-, \pi_k^-)\}$ ). Then, the classical particle filtering algorithm, so-called Condensation algorithm, can be summarized as follows [34].

1. Draw  $K$  particles  $s_k^-$  from the probability density that is represented by the collection  $\{(s_k^-, \pi_k^-)\}$ .
2. Propagate each particle  $s_k^-$  with the transition probability  $p(\alpha|\alpha^-)$  in order to arrive at a collection of  $K$  particles  $s_k$ .
3. Compute the weights  $\pi_k$  for each particle as  $\pi_k = p(y|s_k)$  and then normalize so that  $\sum_k \pi_k = 1$ .

This results in a collection of  $K$  particles and their corresponding weights (i.e.  $\{(s_k, \pi_k)\}$ ), which is an approximation of the density  $p(\alpha|Y)$ .

The Condensation algorithm has three major drawbacks. The first drawback is that a large amount of particles that result from sampling from the proposal density  $p(\alpha|Y^-)$  might be wasted because they are propagated into areas with small likelihood. The second problem is that the scheme ignores the fact that while a particle  $s_k = \langle s_{k1}, s_{k2}, \dots, s_{kN} \rangle$  might have low likelihood, it can easily happen that parts of it might be close to the correct solution. Finally, the third problem is that the estimation of the particle weights does not take into account the interdependences between the different parts of the state  $\alpha$ .

To track facial feature points for the purposes of facial expression analysis, we utilize two different extensions to classical Condensation algorithm. The first one is the Auxiliary Particle Filtering introduced by Pitt and Shepard [62], which addresses the first drawback of the Condensation algorithm by favoring particles that end up in areas with high likelihood when propagated with the transition density  $p(\alpha | \alpha^-)$ . The second extension to classical Condensation algorithm that we utilize for facial point tracking is the Particle Filtering with Factorized Likelihoods proposed by Patras and Pantic [57]. This algorithm addresses all of the aforementioned problems inherent in the Condensation algorithm by extending the Auxiliary Particle Filtering to take into account the interdependencies between the different parts of the state  $\alpha$ . In order to do so we partition the state  $\alpha$  into sub-states  $\alpha_i$  that correspond to the different facial features, that is  $\alpha = \langle \alpha_1, \dots, \alpha_n \rangle$ . At each frame of the sequence we obtain a particle-based representation of  $p(\alpha | y)$  in two stages. In the first stage, we apply one complete step of a particle filtering algorithm (in our case the auxiliary particle filtering) in order to obtain a particle-based representation of  $p(\alpha_i | y)$ , for each facial feature  $i$ . That is, at the first stage, each facial feature is tracked for one frame independently from the other facial features. At the second stage, interdependencies between the sub-states are taken into consideration, in a scheme that samples complete particles from the proposal distribution  $\prod_i p(\alpha_i | y)$  and evaluates them using  $p(\alpha | \alpha^-)$ . The density  $p(\alpha | \alpha^-)$ , that captures the interdependencies between the locations of the



**Fig. 8.** Results of the facial point tracking in face-profile image sequences [49]. First row: frames 1 (neutral), 48 (onset AU29), 59 (apex AU29), and 72 (offset AU29). Second row: frames 1 (neutral), 25 (onset AU12), 30 (onset AU6+12), and 55 (apex AU6+12+25+45).

facial features is estimated using a kernel-based density estimation scheme. Finally, we define an observation model that is based on a robust color-based distance between the color template  $o = \{o_i | i = 1 \dots M\}$  and a color template  $c = \{c_i | i = 1 \dots M\}$  at the current frame. We attempt to deal with shadows by compensating for the global intensity changes. We use the distance function  $d$ , defined by equation (3), where  $M$  is the number of pixels in each template,  $m_c$  is the average intensity of template  $c = \{c_i\}$ ,  $m_o$  is the average intensity of template  $o = \{o_i\}$ ,  $i$  is the pixel index, and  $\rho(\cdot)$  is a robust error function such as the Geman-McClure.

$$d = \sum_{i=1}^M \rho \left( \left\| \frac{c_i}{m_c} - \frac{o_i}{m_o} \right\|_1 \mu_c \right) / M . \quad (3)$$

Typical results of the Auxiliary Particle Filtering, adapted for the problem of color-based template tracking as explained above and applied for tracking facial points in video sequences of profile view of the face [49], are shown in Fig. 8. Typical results of the Particle Filtering with Factorized Likelihoods, applied for tracking color-based templates of facial points in image sequences of faces in frontal-view [81], are shown in Fig. 9.



**Fig. 9.** Results of the facial point tracking in frontal-view face image sequences [81]. First row: frames 1 (neutral), 14 (onset AU1+2+5+20+25+26), and 29 (apex AU1+2+5+20+25+26). Second row: frames 1 (neutral), 32 (apex AU4+7+17+24), and 55 (apex AU45, offset AU4+7+17+24).

## 5 Facial Action Coding

Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, happiness, anger, and surprise (e.g. [3,10,14,32,42,43,89]; for exhaustive surveys of the past work on this research topic, the reader is referred to [51,52,68]). This practice may follow from the work of Darwin [18] and more recently Ekman [22,24,38], who suggested that basic emotions have corresponding prototypic facial expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features, such as the raising of the eyebrows in surprise [46]. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in various applications summarized in Sections 1.1 and 1.2, automatic recognition of rapid facial signals, i.e., facial muscle actions, such as the action units (AUs) of the Facial Action Coding System (FACS) [25,26], is needed.

### 5.1 Facial Action Coding System























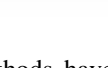
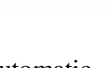
Rapid facial signals are movements of the facial muscles that pull the skin, causing a temporary distortion of the shape of the facial features and of the appearance of folds, furrows, and bulges of skin. The common terminology for describing rapid facial signals refers either to culturally dependent linguistic terms indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown, sneer) or to the linguistic universals describing the activity of specific facial muscles that caused the observed facial appearance changes.

There are several methods for linguistically universal recognition of facial changes based on the facial muscular activity [69]. From those, the facial action coding system (FACS) proposed by Ekman et al. [25,26] is the best known and most commonly used system. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. The changes in the facial expression are described with FACS in terms of 44 different Action Units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or a set of facial muscles. Examples of different AUs are given in Table 2. Along with the definition of various AUs, FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset) in a face image. Using these rules, a FACS coder (that is a human expert having a formal training in using FACS) decomposes a shown facial expression into the AUs that produce the expression.

### 5.2 Automated Facial Action Coding

Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by a computer is by no means a trivial task. A problematic issue is that AUs can occur in more than 7000 different complex combinations [69], causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-of-image-plane movements of permanent facial features (e.g., jetted jaw) that are difficult to detect in 2D face images.

**Table 2.** Examples of Facial Action Units (AUs) defined by the FACS system [25,26]

|   |  |   |  |
|---|--|---|--|
|    | AU1:<br>Raised inner eyebrow                       |    | AU2:<br>Raised outer eyebrow                       |
|    | AU1 + AU2:<br>Raised eyebrows                      |    | AU4:<br>Lowered eyebrow<br>Eyebrows drawn together |
|    | AU5:<br>Raised upper eyelid                        |    | AU6:<br>Raised cheek<br>Compressed eyelid          |
|    | AU7:<br>Tightened eyelid                           |    | AU41:<br>Drooped eyelid                            |
|    | AU44:<br>Squinted eyes                             |    | AU46:<br>Wink                                      |
|    | AU9:<br>Wrinkled nose                              |    | AU11:<br>Deepened nasolabial furrow                |
|    | AU12:<br>Lip corners pulled up                     |    | AU13:<br>Lip corners pulled up<br>sharply          |
|    | AU14:<br>Dimpler - mouth<br>corners pulled inwards |    | AU15:<br>Lip corners depressed                     |
|    | AU17:<br>Chin raised                               |    | AU19:<br>Tongue shown                              |
|    | AU20:<br>Mouth stretched<br>horizontally           |    | AU24:<br>Lips pressed                              |
|   | AU26:<br>Jaw dropped                               |   | AU29:<br>Jaw pushed forward                        |
|  | AU30:<br>Jaw sideways                              |  | AU36:<br>Bulge produced by the<br>tongue           |

Few methods have been reported for automatic AU detection in face image sequences [37,51,78]. Some researchers described patterns of facial motion that correspond to a few specific AUs, but did not report on actual recognition of these AUs (e.g. [10,29,42,43,76]). Only recently there has been an emergence of efforts toward explicit automatic analysis of facial expressions into elementary AUs [47,78]. For instance, the Machine Perception group at UCSD has proposed several methods for automatic AU coding of input face video. To detect 6 individual AUs in face image sequences free of head motions, Bartlett et al. [7] used a  $61 \times 10 \times 6$  feed-forward neural network. They achieved 91% accuracy by feeding the pertinent network with the results of a hybrid system combining holistic spatial analysis and optical flow with local feature analysis. To recognize 8 individual AUs and 4 combinations of AUs in face image sequences free of head motions, Donato et al. [21] used Gabor wavelet representation and independent component analysis. They reported a 95.5% average recognition rate achieved by their method. The most recent work by Bartlett et al. [8] reports on accurate automatic recognition of 18 AUs (95% average recognition rate)



from near frontal-view face image sequences using Gabor wavelet features and a classification technique based on AdaBoost and Support Vector Machines (SVM). Another group that has focused on automatic FACS coding of face image sequences is the CMU group led by Takeo Kanade and Jeff Cohn. To recognize 8 individual AUs and 7 combinations of AUs in face image sequences free of head motions, Cohn et al. [16] used facial feature point tracking and discriminant function analysis and achieved an 85% average recognition rate. Tian et al. [77] used lip tracking, template matching and neural networks to recognize 16 AUs occurring alone or in combination in near frontal-view face image sequences. They reported an 87.9% average recognition rate.

Our group also reported on multiple efforts toward automatic analysis of facial expressions into atomic facial actions. The majority of this previous work concerns automatic AU recognition in static face images [50,53]. To our best knowledge, these systems are the first (and at this moment the only) to handle AU detection in static face images. However, these works are not relevant to the present goals, since they cannot handle video streams inherent in AmI applications. Only recently, our group has focused on automatic FACS coding of face video. To recognize 15 AUs occurring alone or in combination in near frontal-view face image sequences, Valstar et al. [80] used temporal templates (Section 3.2) and compared two classification techniques: (i) a combined k-Nearest-Neighbor and rule-based classifier, and (ii) a SNoW classifier. An average recognition rate ranging from 56% to 68% has been achieved. Except for this work, and based upon the tracked movements of facial characteristic points, we mainly experimented with rule-based [48,49] and SVM-based methods [81] for recognition of AUs in either near frontal-view (Fig. 10) or near profile-view (Fig. 11) face image sequences.

A basic understanding of how to achieve automatic AU detection from the profile view of the face is necessary if a technological framework for automatic AU detection from multiple views of the face is to be established [49]. The automatic AU detection from the profile view of the face was deemed the most promising method for achieving robust AU detection [86], independent of rigid head movements that can cause changes in the viewing angle and the visibility of the tracked face and its features. To our knowledge, our system for AU recognition from face profile-view image sequences is the first (and at this moment the only) to address this problem.

In contrast to the aforementioned methods developed elsewhere, which address mainly the problem of spatial modeling of facial expressions, the methods proposed by our group address the problem of temporal modeling of facial expressions as well. In other words, the methods proposed here are very suitable for encoding temporal activation patterns (onset  $\rightarrow$  apex  $\rightarrow$  offset) of AUs shown in an input face video. This is of importance for there is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., the timing and the duration of facial activity) is a critical factor for the interpretation of the observed behavior [67]. For example, Schmidt and Cohn [70] have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1 second. Since it takes more than

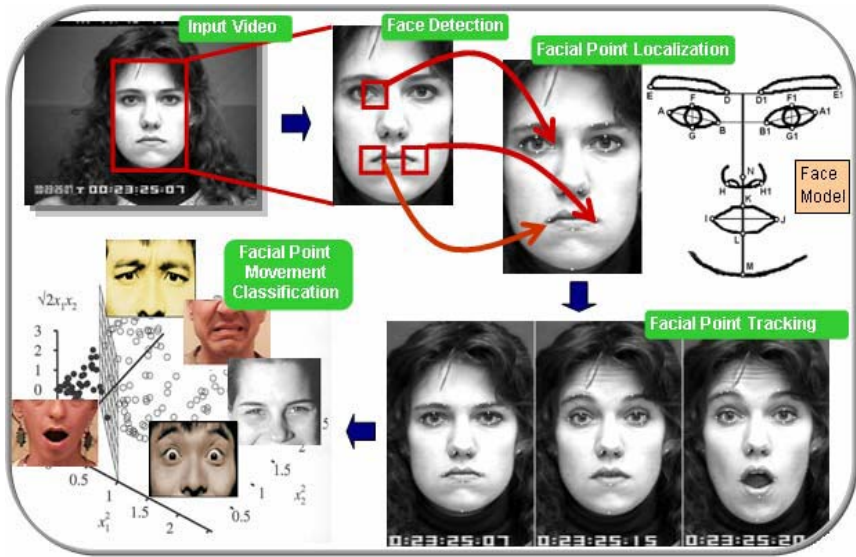


Fig. 10. Outline of our AU detectors from frontal-view face image sequences [48,81]

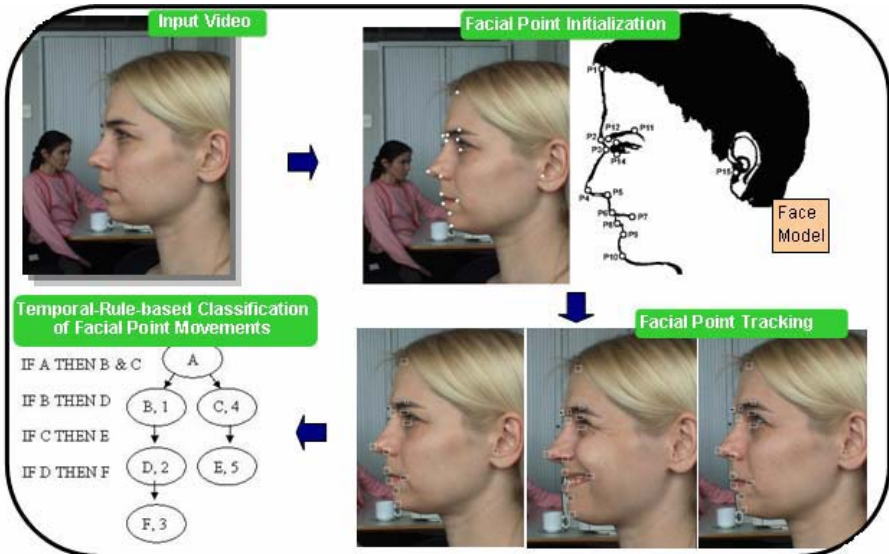


Fig. 11. Outline of our AU detector from profile-view face image sequences [49]

one hour to manually score 100 still images or a minute of videotape in terms of AUs and their temporal segments [25], it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. To our best knowl-

edge, our systems are the first (and at this moment the only) to explicitly handle temporal segments of AUs.

To recognize a set of 27 AUs occurring alone or in combination in a near frontal-view face image sequence [48], we proceed under 2 assumptions (as defined for video samples of either the Cohn-Kanade [37] or the MMI facial expression database [56]): (1) the input image sequence is non-occluded near frontal-view of the face, and (2) the first frame shows a neutral expression and no head rotations. To handle possible in-image-plane head rotations and variations in scale of the observed face, we register each frame of the input image sequence with the first frame based on three referential points (Fig. 10): the tip of the nose (N) and the inner corners of the eyes (B and B1). We use these points as the referential points because of their stability with respect to non-rigid facial movements: facial muscle actions do not cause physical displacements of these points. Each frame is registered with the first frame by applying an affine transformation. Except of N, B and B1, which are tracked in unregistered input video sequences, other facial fiducial points are tracked in the registered input image sequence. Typical tracking results are shown in Fig. 9. Based upon the changes in the position of the fiducial points, we measure changes in facial expression. Changes in the position of the fiducial points are transformed first into a set of mid-level parameters for AU recognition. We defined two parameters: *up/down(P)* and *inc/dec(PP')*. Parameter *up/down(P)* =  $y(P_{t1}) - y(P_t)$  describes upward and downward movements of point  $P$  and parameter *inc/dec(PP')* =  $PP'_{t1} - PP'_t$  describes the increase or decrease of the distance between points  $P$  and  $P'$ . Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination in the input face videos. For instance, to recognize the temporal segments of AU12 (Table 2), which pulls the mouth corners upwards in a smile, we exploit the following temporal rules (for experiments with a SVM-based binary classifier instead of rules, see [81]):

```

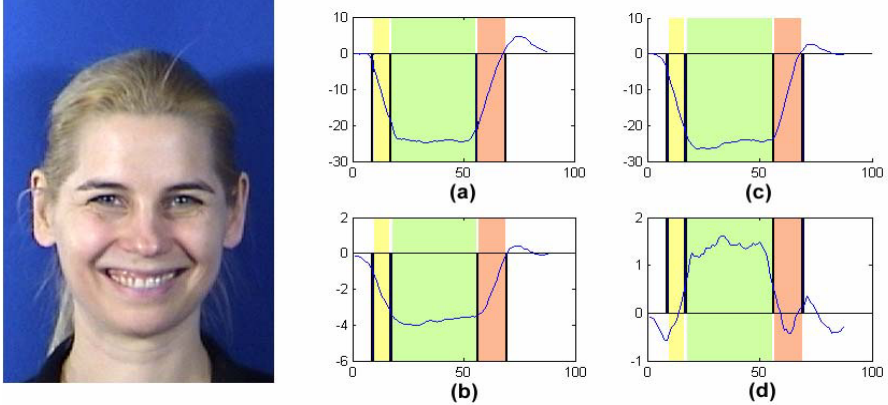
IF (up/down(I) >  $\epsilon$  AND inc/dec(NI) <  $-\epsilon$ )
  OR (up/down(J) >  $\epsilon$  AND inc/dec(NJ) <  $-\epsilon$ ) THEN AU12-p
IF AU12-p AND { (up/down(I))t > [up/down(I)]t-1 )
  OR ( [up/down(J)]t > [up/down(J)]t-1 ) } THEN AU12-onset
IF AU12-p AND { ( | [up/down(I)]t - [up/down(I)]t-1 | ≤  $\epsilon$  )
  OR ( | [up/down(J)]t - [up/down(J)]t-1 | ≤  $\epsilon$  ) } THEN AU12-apex
IF AU12-p AND { (up/down(I))t < [up/down(I)]t-1 )
  OR ( [up/down(J)]t < [up/down(J)]t-1 ) } THEN AU12-offset

```

Fig. 12 illustrates the meaning of these rules. The horizontal axis represents the time dimension (i.e., the frame number) and the vertical axis represents values that the mid-level feature parameters take. As implicitly suggested by the graphs of Fig. 12, I and/or J should move upward and be above their neutral-expression location to label a frame as an “AU12<sup>2</sup> onset”. The upward motion should terminate, resulting in a (relatively) stable temporal location of I and/or J, for a frame to be labeled as “AU12

<sup>2</sup> Since the upward motion of the mouth corners is the principle cue for the activation of AU12, the upward movement of the fiducial points I and/or J (i.e., point P7 in the case of the profile view of the face) is used as the criterion for detecting the onset of the AU12 activation. Reversal of this motion is used to detect the offset of this facial expression.

apex”. Eventually, I and/or J should move downward toward their neutral-expression location to label a frame as an “AU12 offset”. Note that, at the end of the offset phase, the graphs show a distinct increase of the values of the mid-level parameters, beyond their neutral-expression values. As shown by Schmidt and Cohn [70], this is typical for so-called “dampened” spontaneous smiles and in contrast to posed smiles.



**Fig. 12.** The changes in x and y coordinates of points I and J (mouth corners) computed for 90 frames of an AU6+12+25 frontal-view face video (the apex frame is illustrated). (a)-(b) Point I. (c)-(d) Point J.

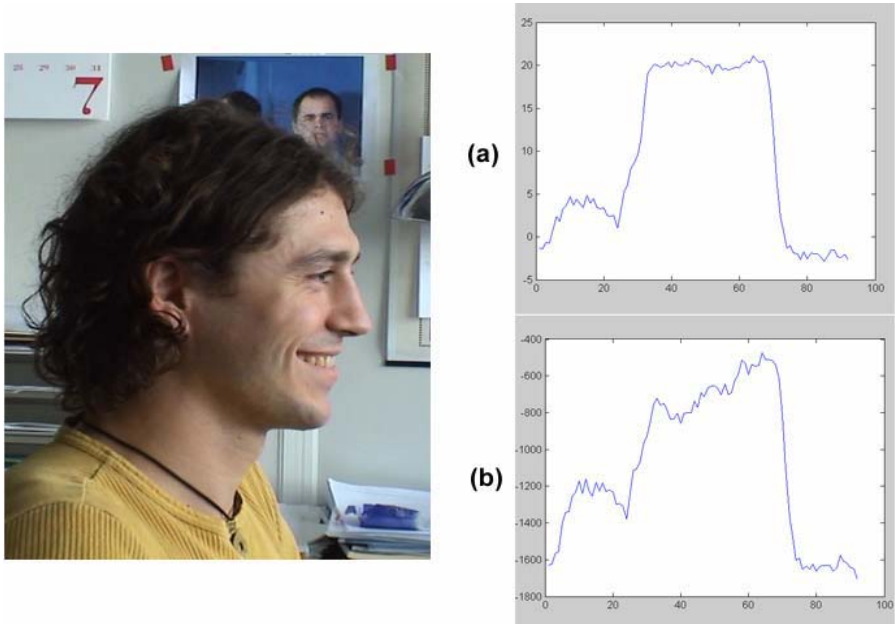
Similarly, to recognize a set of 27 AUs occurring alone, or in combination in a near profile-view face image sequence [49], we proceed under 2 assumptions (as defined for video samples of the MMI facial expression database [56]): (1) the input image sequence is non-occluded (left or right) near profile-view of the face with possible in-image-plane head rotations, and (2) the first frame shows a neutral expression. To make the processing robust to in-image-plane head rotations and translations as well as to small translations along the z-axis, we estimate a global affine transformation  $\vartheta$  for each frame, and based on it we register the current frame to the first frame of the sequence. In order to estimate the global affine transformation, we track three referential points. These are (Fig. 11): the top of the forehead (P1), the tip of the nose (P4), and the ear canal entrance (P15). We use these points as the referential points because of their stability with respect to non-rigid facial movements. We estimate the global affine transformation  $\vartheta$  as the one that minimizes the distance (in the least-squares sense) between the  $\vartheta$ -based projection of the tracked locations of the referential points and these locations in the first frame of the sequence. The rest of the facial points illustrated in Fig. 11 are tracked in frames that have been compensated for the transformation  $\vartheta$ . Typical tracking results are shown in Fig. 8. Changes in the position of the facial points are transformed first into a set of mid-level parameters for AU recognition described above. Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination in the input face videos. For instance, to recognize the temporal segments of AU12, we exploit the following temporal rules:

```

IF ( $up/down(P7) > \epsilon$  AND  $inc/dec(P5P7) \geq \epsilon$ ) THEN AU12-p
IF AU12-p AND ( $([up/down(P7)]_t > [up/down(P7)]_{t-1})$ ) THEN
AU12-onset
IF AU12-p AND ( $(|[up/down(P7)]_t - [up/down(P7)]_{t-1}| \leq \epsilon)$ )
THEN AU12-apex
IF AU12-p AND ( $([up/down(P7)]_t < [up/down(P7)]_{t-1})$ ) THEN
AU12-offset

```

Fig. 13 illustrates the meaning of these rules. P7 should move upward, above its neutral-expression location, and the distance between points P5 and P7 should increase, exceeding its neutral-expression length, in order to label a frame as an “AU12 onset”. In order to label a frame as “AU12 apex”, the increase of the values of the relevant mid-level parameters should terminate. Once the values of these mid-level parameters begin to decrease, a frame can be labeled as “AU12 offset”. Note that the graphs of Fig. 13 show two distinct peaks in the increase of the pertinent mid-level parameters. According to [70], this is typical for spontaneous smiles.



**Fig. 13.** The values of mid-level parameters (a)  $up/down(P7)$  and (b)  $inc/dec(P5P7)$  computed for 92 frames of AU6+12+25 face-profile video (the apex frame is illustrated)

We tested our method for AU coding in near frontal-view face image sequences on both Cohn-Kanade [37] and MMI facial expression database [56]. The accuracy of the method was measured with respect to the misclassification rate of each “expressive” segment of the input sequence, not with respect to each frame [48]. Overall, for 135 test samples from both databases, we achieved an average recognition rate of 90% sample-wise for 27 different AUs occurring alone or in combination in an input video.

Since Cohn-Kanade database does not contain images of faces in profile view (it contains only displays of emotions recorded in frontal facial view), the method for AU coding in near profile-view face video was tested on MMI facial expression database only. The accuracy of the method was measured with respect to the misclassification rate of each “expressive” segment of the input sequence[49]. Overall, for 96 test samples, we achieved an average recognition rate of 87% sample-wise for 27 different AUs occurring alone, or in combination, in an input video.

## 6 Facial Expression Interpretation and Facial Affect Recognition

As already noted above, virtually all systems for automatic facial expression analysis attempt to recognize a small set of universal/basic emotions [51,52]. However, pure expressions of “basic” emotions are seldom elicited; most of the time people show blends of emotional displays [38]. Hence, the classification of facial expressions into a single “basic”-emotion category is not realistic. Also, not all facial actions can be classified as a combination of the “basic” emotion categories. Think, for instance, about fatigue, frustration, anxiety, or boredom. In addition, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state may differ from culture to culture and even from person to person [67]. Furthermore, not all facial actions should be associated with affective states. Think, for instance, about face-based interface-control systems for support of disabled users. Hence, pragmatic choices (user- and use-case-profiled choices) must be made regarding the selection of interpretation labels (such as affective labels) to be assigned by an automatic system to sensed facial signals. This is especially the case with AmI technologies, where one basic goal is to ensure that products are tailored to the user’s preferences, needs and abilities (Table 1).

We developed a case-based reasoning system that learns its expertise by having the user instruct the system on the desired (context-sensitive) interpretations of sensed facial signals [54]. To our best knowledge, it is the first (and at this moment the only) system facilitating user-profiled interpretation of facial expressions. We used it to achieve classification of AUs into the emotion categories learned from the user.

Since AUs can occur in more than 7000 combinations [69], the classification of AUs in an arbitrary number of emotion categories learned from the user is a very complex problem. To tackle this problem, one can apply either eager or lazy learning methods. Eager learning methods, such as neural networks, extract as much information as possible from training data and construct a general approximation of the target function. Lazy learning methods, such as case-based reasoning, simply store the presented data and generalizing beyond these data is postponed until an explicit request is made. When a query instance is encountered, similar related instances are retrieved from the memory and used to classify the new instance. Hence, lazy methods have the option of selecting a different local approximation of the target function for each presented query instance. Eager methods using the same hypothesis space are more restricted because they must choose their approximation before presented queries are observed. In turn, lazy methods are usually more appropriate for complex and incomplete problem domains than eager methods, which replace the training data with abstractions obtained by generalization and which, in turn, require an excessive amount

of training data. Therefore, we chose to achieve classification of the AUs detected in an input face video into the emotion categories learned from the user by case-based reasoning, a typical lazy learning method.

The utilized case base is a dynamic, incrementally self-organizing event-content-addressable memory that allows fact retrieval and evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Each event (case) is one or more micro-events, each of which is a set of AUs. Micro-events related by the goal of communicating one specific affective state are grouped within the same dynamic memory chunk. In other words, each memory chunk represents a specific emotion category and contains all micro-events to which the user assigned the emotion label in question. The indexes associated with each dynamic memory chunk comprise individual AUs and AU combinations that are most characteristic for the emotion category in question. Finally, the micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of times a given micro-event occurred, the higher its hierarchical position within the given chunk. The initial endowment of the dynamic memory is achieved by asking the user to associate an interpretation (emotion) label to a set of 40 typical facial expressions (micro-events that might be hardwired to emotions according to [46]). Fig. 14 illustrates a number of examples of the utilized stimulus material.



**Fig. 14.** Sample stimulus images from the MMI Facial Expression Database [56] used for initial endowment of the case base. Left to right: AU1+2, AU10, AU6+12, AU15+17.

The classification of the detected AUs into the emotion categories learned from the user is further accomplished by case-based reasoning about the content of the dynamic memory. To solve a new problem of classifying a set of input AUs into the user-defined interpretation categories, the following steps are taken:

1. Search the dynamic memory for similar cases, retrieve them, and interpret the input set of AUs using the interpretations suggested by the retrieved cases.
2. If the user is satisfied with the given interpretation, store the case in the dynamic memory. Otherwise, adapt the memory according to user-provided feedback on the interpretation he associates with the input facial expression.

The utilized retrieval and adaptation algorithms employ a pre-selection of cases that is based upon the clustered organization of the dynamic memory, the indexing structure

of the memory, and the hierarchical organization of cases within the clusters/ chunks according to their typicality [54].

Two validation studies on a prototype system have been carried out. The question addressed by the 1<sup>st</sup> validation study was: How acceptable are the interpretations given by the system after it is trained to recognize 6 basic emotions? The question addressed by the 2<sup>nd</sup> validation study was: How acceptable are the interpretations given by the system, after it is trained to recognize an arbitrary number of user-defined interpretation categories? In the first case, a human FACS coder was asked to train the system. In the second case, a lay expert, without formal training in emotion signals recognition, was asked to train the system. The same expert used to train the system was used to evaluate its performance, i.e., to judge the acceptability of interpretations returned by the system. For basic emotions, in 100% of test cases the expert approved of the interpretations generated by the system. For user-defined interpretation categories, in 83% of test cases the lay expert approved entirely of the interpretations and in 14% of test cases the expert approved of most but not of all the interpretation labels generated by the system for the pertinent cases. These findings indicate that the facial expression interpretation achieved by the system is rather accurate.

## 7 Conclusions

Automating the analysis of facial signals, especially rapid facial signals (i.e., AUs), is important to realize context-sensitive, face-based (multimodal) ambient interfaces, to advance studies on human emotion and affective computing, and to boost numerous applications in fields as diverse as security, medicine, and education. This paper provided an overview of the efforts of our research group in approaching this goal. To summarize:

- Our methods for automatic facial feature point detection and tracking extend the state of the art in facial point detection and tracking in several ways, including the number of facial points detected (20 in total), the accuracy of the achieved detection (93% of the automatically detected points were displaced in any direction, horizontal or vertical, less than 5% of the inter-ocular distance [83]), the accuracy, and the robustness of the tracking scheme including the invariance to noise, occlusion, clutter and changes in the illumination intensity (inherent in Particle Filtering with Factorized Likelihoods [57,58]).
- Our approaches to automatic AU coding of face image sequences extend the state of the art in the field in several ways, including the facial view (profile), the temporal segments of AUs (onset, apex, offset), the number (27 in total), and the difference in AUs (e.g. AU29, AU36) handled. To wit, the automated systems for AU detection from face video that have been reported so far do not deal with the profile view of the face, cannot handle temporal dynamics of AUs, cannot detect out-of-plane movements such as thrusting the jaw forward (AU29), and, at best, can detect 16 to 18 AUs (from in total 44 AUs). The basic insights in how to achieve automatic detection of AUs in profile-face videos and how to realize automatic detection of temporal segments of AUs in either frontal- or profile-view face image sequences can aid and abet further research on facial expression symmetry, spontaneous vs. posed facial expressions, and facial expression recognition from multiple facial views [48,49].



- We also proposed a new facial expression interpretation system that performs classification of AUs into the emotion categories learned from the user [54]. Given that the previously reported systems for high-abstraction-level analysis of facial expressions are able to classify facial expressions only in one of the 6 basic emotion categories, our facial expression interpreter extends the state of the art in the field by enabling facial signal interpretation in a user-adaptive manner. Further research on accomplishing context-sensitive (user-, task-, use-case-, environment-dependent) interpretation of facial (or any other behavioral) signals can be based upon our findings. This is especially important for AmI technologies where one basic goal is to ensure that products are tailored to the user's preferences, needs and abilities.

However, our methods cannot recognize the full range of facial behavior (i.e. all 44 AUs defined in FACS); they detect up to 27 AUs occurring alone or in combination in near frontal- or profile-view face image sequences. A way to deal with this problem is to look at diverse facial features. Although it has been reported that methods based on geometric features are usually outperformed by those based on appearance features using, e.g., Gabor wavelets or eigenfaces [7], our studies have shown that this claim does not always hold [49,81]. We believe, however, that further research efforts toward combining both approaches are necessary if the full range of human facial behavior is to be coded in an automatic way.

If we consider the state of the art in face detection and facial point localization and tracking, then noisy and partial data should be expected. As remarked by Pantic et al. [47,52], a facial expression analyzer should be able to deal with these imperfect data and to generate its conclusion so that the certainty associated with it varies with the certainty of face and facial point localization and tracking data. Our point tracker is very robust to noise, occlusion, clutter and changes in lighting conditions and it deals with inaccuracies in facial point tracking using a memory-based process that takes into account the dynamics of facial expressions [49,57,58]. However, our methods do not calculate the output data certainty by propagating the input data certainty (i.e. the certainty of facial point tracking). Future work on this issue aims at investigating the use of measures that can express the confidence to facial point tracking and that can facilitate both a more robust AU recognition and the assessment of the certainty of the performed AU recognition.

Finally, our methods assume that the input data are near frontal- or profile-view face image sequences showing facial displays that always begin with a neutral state. In reality, such assumption cannot be made; variations in the viewing angle should be expected. Also, human facial behavior is more complex and transitions from a facial display to another do not have to involve intermediate neutral states. Consequently, our facial expression analyzers cannot deal with spontaneously occurring (unposed) facial behavior. In turn, actual deployment of our methods in ambient interfaces and AmI sensing technologies is still in the relatively distant future. There are a number of related issues that should be addressed. How to achieve parsing of the stream of facial and head movements not under volitional control? What properties should automated analyzers of human expressive behavior have in order to be able to analyze human spontaneous behavior? How should one elicit spontaneous human expressive behavior, including genuine emotional responses, necessary for the training automated systems? How should the grammar of human expressive behavior be learned?

Tian et al. [78] and Pantic et al. [52,55] have discussed some of these aspects of automated facial expression analysis and they form the main focus of our current and future research efforts. Yet, since the complexity of these issues concerned with the interpretation of human behavior at a deeper level is tremendous and spans several different disciplines in computer and social sciences, we believe that a large, focused, interdisciplinary, international program directed towards computer understanding and responding to human behavioral patterns (as shown by means of facial expressions and other modes of social interaction) should be established if we are to experience breakthroughs in human-computer and ambient interface designs.

## Acknowledgements

The work of M. Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202.

## References

- [1] Aarts, E.: Ambient Intelligence – Visualizing the Future. Proc. Conf. Smart Objects & Ambient Intelligence (2005) (<http://www.soc-eusai2005.org/>)
- [2] Ambady, N., Rosenthal, R.: Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, Vol. 111, No. 2 (1992) 256-274
- [3] Anderson, K., McOwan, P.W.: A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 36, No. 1 (2006) 96-105
- [4] Aristotle: *Physiognomonica*. In: Ross, W.D. (ed.): *The works of Aristotle*. Clarendon, Oxford (nd/1913) 805-813
- [5] Baker, S., Matthews, I., Xiao, J., Gross, R., Kanade, T.: Real-time non-rigid driver head tracking for driver mental state estimation. Proc. World Congress on Intelligent Transportation Systems (2004) ([http://www.ri.cmu.edu/projects/project\\_448.html](http://www.ri.cmu.edu/projects/project_448.html))
- [6] Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. *J. Computer Vision*, Vol. 12, No. 1 (1994) 43-78
- [7] Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Measuring facial expressions by computer image analysis. *Psychophysiology*, Vol. 36 (1999) 253-263
- [8] Bartlett, M.S., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.R.: Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. Proc. Conf. Systems, Man, and Cybernetics, Vol. 1 (2004) 592-597.
- [9] Bassili, J.N.: Facial Motion in the Perception of Faces and of Emotional Expression. *J. Experimental Psychology*, Vol. 4 (1978) 373-379
- [10] Black, M., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. *Computer Vision*, Vol. 25, No. 1 (1997) 23-48
- [11] Bobick, A.F., Davis, J.W.: The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3 (2001) 257-267
- [12] Bowyer, K.W.: Face Recognition Technology – Security vs. Privacy. *IEEE Technology and Society Magazine*, Vol. 23, No. 1 (2004) 9-19
- [13] Bruce, V.: *Recognizing Faces*. Lawrence Erlbaum Assoc., Hove (1986)

- [14] Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences – temporal and static modeling. *Computer Vision and Image Understanding*, Vol. 91 (2003) 160-187
- [15] Cohn, J.F., Reed, L.I., Ambadar, Z., Xiao, J., Moriyma, T.: Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. *Proc. Conf. Systems, Man and Cybernetics*, Vol. 1 (2004) 610-616
- [16] Cohn, J.F., Zlochower, A.J., Lien, J., Kanade, T.: Automated face analysis by feature point tracking has high concurrent validity with manual faces coding, *Psychophysiology*, Vol. 36 (1999) 35-43
- [17] Cristinacce, D., Cootes, T.F.: A Comparison of Shape Constrained Facial Feature Detectors. *Proc. Conf. Automatic Face and Gesture Recognition (2004)* 375-380
- [18] Darwin, C.: *The expression of the emotions in man and animals*. University of Chicago Press, Chicago (1965 / 1872)
- [19] DeCarlo, D., Metaxas, D.: The integration of optical flow and deformable models with applications to human face shape and motion estimation. *Proc. Conf. Computer Vision and Pattern Recognition (1996)* 231-238
- [20] Dishman, E.: Inventing wellness systems for aging in place. *IEEE Computer Magazine, Spec. Issue on Computing and the Aging*, Vol. 37, No. 5 (2004) 34-41
- [21] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10 (1999) 974-989
- [22] Ekman, P.: *Emotions Revealed*. Times Books, New York (2003)
- [23] Ekman, P., Friesen, W.V.: *The repertoire of nonverbal behavior*. *Semiotica*, Vol. 1 (1969) 49-98
- [24] Ekman, P., Friesen, W.V.: *Unmasking the face*. Prentice-Hall, New Jersey (1975)
- [25] Ekman, P., Friesen, W.V.: *Facial Action Coding System*. Consulting Psychologist Press, Palo Alto (1978)
- [26] Ekman, P., Friesen, W.V., Hager, J.C.: *Facial Action Coding System. A Human Face*, Salt Lake City (2002)
- [27] Fasel, I., Fortenberry, B., Movellan, J.R.: GBoost: A generative framework for boosting with applications to real-time eye coding. *Computer Vision and Image Understanding*, under review (<http://mplab.ucsd.edu/publications/>)
- [28] Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, Vol. 28, No. 2 (2000) 337-374
- [29] Gokturk, S.B., Bouguet, J.Y., Tomasi, C., Girod, B.: Model-based face tracking for view-independent facial expression recognition. *Proc. Conf. Automatic Face and Gesture Recognition (2002)* 272-278.
- [30] Gross, T.: Ambient Interfaces – Design Challenges and Recommendations. *Proc. Conf. Human-Computer Interaction (2003)* 68-72
- [31] Gu, H., Ji, Q.: Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, Vol. 16, No. 2 (2005) 105-115
- [32] Guo, G., Dyer, C.R.: Learning From Examples in the Small Sample Case – Face Expression Recognition. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 35, No. 3 (2005) 477-488
- [33] Haykin, S., de Freitas, N. (eds.): *Special Issue on Sequential State Estimation*. *Proceedings of the IEEE*, vol. 92, No. 3 (2004) 399-574
- [34] Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. *J. Computer Vision*, Vol. 29, No. 1 (1998) 5-28

- [35] Jacobs, D.W., Osadchy, M., Lindenbaum, M.: What Makes Gabor Jets Illumination In-sensitive? (<http://rita.osadchy.net/papers/gabor-3.pdf>)
- [36] Kalman, R.E.: A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.*, Vol. 82 (1960) 35-45
- [37] Kanade, T., Cohn, J., Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proc. Conf. Automatic Face and Gesture Recognition (2000)* 46-53.
- [38] Keltner, D., Ekman, P.: Facial Expression of Emotion. In: Lewis, M., Haviland-Jones, J.M. (eds.): *Handbook of Emotions*. 2<sup>nd</sup> edition. The Guilford Press, New York (2004) 236-249
- [39] Li, S.Z., Jain, A.K. (eds.): *Handbook of Face Recognition*. Springer, New York (2005)
- [40] van Loenen, E.J.: On the role of Graspable Objects in the Ambient Intelligence Paradigm. *Proc. Conf. Smart Objects (2003)* (<http://www.grenoble-soc.com/>)
- [41] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. *Proc. Conf. Artificial Intelligence (1981)* 674-679
- [42] Martinez, A.M.: Matching expression variant faces. *Vision Research*, Vol. 43 (2003) 1047-1060
- [43] Mase, K.: Recognition of facial expression from optical flow. *IEICE Transactions*, Vol. E74, No. 10 (1991) 3474-3483
- [44] Moghaddam, B., Pentland, A.: Probabilistic Visual Learning for Object Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7 (1997) 696-710
- [45] Norman, D.A.: *The Invisible Computer*. MIT Press, Cambridge (1999)
- [46] Ortony, A., Turner, T.J.: What is basic about basic emotions? *Psychological Review*, Vol. 74 (1990) 315-341
- [47] Pantic, M.: Face for Interface. In: Pagani, M. (ed.): *The Encyclopedia of Multimedia Technology and Networking 1*. Idea Group Reference, Hershy (2005) 308-314
- [48] Pantic, M., Patras, I.: Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. Conf. Systems, Man, and Cybernetics (2005)*
- [49] Pantic, M., Patras, I.: Dynamics of Facial Expressions – Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 36 (2006)
- [50] Pantic, M., Rothkrantz, L.J.M.: Expert system for automatic analysis of facial expression. *Image and Vision Computing*, Vol. 18, No. 11 (2000) 881-905
- [51] Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions – The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12 (2000) 1424-1445
- [52] Pantic, M., Rothkrantz, L.J.M.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE, Spec. Issue on Human-Computer Multimodal Interface*, Vol. 91, No. 9 (2003) 1370-1390
- [53] Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis from Static Face Images, *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 34, No. 3 (2004) 1449-1461
- [54] Pantic, M., Rothkrantz, L.J.M.: Case-based reasoning for user-profiled recognition of emotions from face images, *Proc. Conf. Multimedia and Expo*, Vol. 1 (2005) 391-394
- [55] Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective Multimodal Human-Computer Interaction, *Proc. ACM Conf. Multimedia (2005)*
- [56] Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-based database for facial expression analysis, *Proc. Conf. Multimedia and Expo (2005)* (<http://www.mmifacedb.com/>)

- [57] Patras, I., Pantic, M.: Particle Filtering with Factorized Likelihoods for Tracking Facial Features. Proc. Conf. Automatic Face and Gesture Recognition (2004) 97-102
- [58] Patras, I., Pantic, M.: Tracking Deformable Motion. Proc. Conf. Systems, Man, and Cybernetics (2005)
- [59] Pentland, A.: Looking at people – Sensing for ubiquitous and wearable computing. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 1 (2000) 107-119
- [60] Pentland, A., Moghaddam, B., Starner, T.: View-Based and Modular Eigenspaces for Face Recognition. Proc. Conf. Computer Vision and Pattern Recognition (1994) 84-91
- [61] Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
- [62] Pitt, M.K., Shephard, N.: Filtering via simulation: auxiliary particle filtering. J. Amer. Stat. Assoc., Vol. 94 (1999) 590-599
- [63] Preece, J., Rogers, Y., Sharp, H.: Interaction Design – Beyond Human-Computer Interaction. John Wiley & Sons, New York (2002)
- [64] Raisinghani, M.S., Benoit, A., Ding, J., Gomez, M., Gupta, K., Gusila, V., Power, D., Schmedding, O.: Ambient Intelligence – Changing Forms of Human-Computer Interaction and their Social Implications. J. Digital Information, Vol. 5, No. 4 (2004) 1-8
- [65] Remagnino, P., Foresti, G.L.: Ambient Intelligence – A New Multidisciplinary Paradigm. IEEE Trans. Systems, Man, and Cybernetics, Part A, Spec. Issue on Ambient Intelligence, Vol. 35, No. 1 (2005) 1-6
- [66] Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1 (1998) 23-38
- [67] Russell, J.A., Fernandez-Dols, J.M. (eds.): The Psychology of Facial Expression. Cambridge University Press, Cambridge (1997)
- [68] Samal, A., Iyengar, P.A.: Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition, Vol. 25, No. 1 (1992) 65-77
- [69] Scherer, K.R., Ekman, P. (eds.): Handbook of methods in non-verbal behavior research. Cambridge University Press, Cambridge (1982)
- [70] Schmidt, K.L., Cohn, J.F.: Dynamics of facial expression: Normative characteristics and individual differences. Proc. Conf. Multimedia and Expo (2001) 547-550
- [71] Shadbolt, N.: Ambient Intelligence. IEEE Intelligent Systems, Vol. 18, No. 4 (2003) 2-3
- [72] Shi, J., Tomasi, C.: Good features to track. Proc. Conf. Computer Vision and Pattern Recognition (1994) 593-600
- [73] Stephanidis, C., Akoumianakis, D., Sfyarakis, M., Paramythis, A.: Universal accessibility in HCI. Proc. ERCIM Workshop. User Interfaces For All (1998) (<http://ui4all.ics.forth.gr/UI4ALL-98/proceedings.html>)
- [74] Streitz, N., Nixon, P.: The Disappearing Computer. ACM Communications, Spec. Issue on The Disappearing Computer, Vol. 48, No. 3 (2005) 33-35
- [75] Sung, K.K., Poggio, T.: Example-Based Learning for View-Based Human Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1 (1998) 39-51
- [76] Tao, H., Huang, T.S.: Connected vibrations – a model analysis approach to non-rigid motion tracking. Proc. Conf. Computer Vision and Pattern Recognition (1998) 735-740
- [77] Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 23, No. 2 (2001) 97-115
- [78] Tian, Y.L., Kanade, T., Cohn, J.F.: Facial Expression Analysis. In: Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005)
- [79] Tscheligi, M.: Ambient Intelligence – The Next Generation of User Centeredness. ACM Interactions, Spec. Issue on Ambient Intelligence, Vol. 12, No. 4 (2005) 20-21
- [80] Valstar, M., Pantic, M., Patras, I.: Motion History for Facial Action Detection from Face Video. Proc. Conf. Systems, Man and Cybernetics, Vol. 1 (2004) 635-640

- [81] Valstar, M., Patras, I., Pantic, M.: Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data. Proc. Conf. Computer Vision and Pattern Recognition (2005)
- [82] Viola, P., Jones, M.: Robust real-time object detection. Proc. Int'l Conf. Computer Vision, Workshop on Statistical and Computation Theories of Vision (2001)
- [83] Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using Gabor feature based boosted classifiers. Proc. Conf. Systems, Man and Cybernetics (2005)
- [84] Weiser, M.: The world is not a desktop. ACM Interactions, Vol. 1, No. 1 (1994) 7-8
- [85] Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time Combined 2D+3D Active Appearance Models. Proc. Conf. Computer Vision and Pattern Recognition, Vol. 2 (2004) 535-542
- [86] Yacoob, Y., Davis, L., Black, M., Gavrila, D., Horprasert, T., Morimoto, C.: Looking at People in Action. In: Cipolla, R., Pentland, A. (eds.): Computer Vision for Human-Machine Interaction. Cambridge University Press, Cambridge (1998) 171-187
- [87] Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 1 (2002) 34-58
- [88] Zhai, S., Bellotti, V.: Introduction to Sensing-Based Interaction. ACM Trans. Computer-Human Interaction, Spec. Issue on Sensing-Based Interaction, Vol. 12, No. 1 (2005) 1-2
- [89] Zhang, Y., Ji, Q.: Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequence. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 27, No. 5 (2005) 699-714
- [90] Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition – A literature survey. ACM Computing Surveys, Vol. 35, No. 4 (2003) 399-458

# Empathic Computing

Yang Cai

Carnegie Mellon University  
ycai@cmu.edu

**Abstract.** Empathic computing is an emergent paradigm that enables a system to understand human states and feelings and to share this intimate information. The new paradigm is made possible by the convergence of affordable sensors, embedded processors and wireless ad-hoc networks. The power law for multi-resolution channels and mobile-stationary sensor webs is introduced to resolve the information avalanche problems. As empathic computing is sensor-rich computing, particular models such as semantic differential expressions and inverse physics are discussed. A case study of a wearable sensor network for detection of a falling event is presented. It is found that the location of the wearable sensor is sensitive to the results. From the machine learning algorithm, the accuracy reaches up to 90% from 21 simulated trials. Empathic computing is not limited to healthcare. It can also be applied to solve other everyday-life problems such as management of emails and stress.

## 1 Introduction

The function of aging is disability. Cataracts, Alzheimer and Osteoporosis, for example, are common symptoms of old age. The Impressionist painter Claude Monet is believed to have developed cataracts in later life, and the effect may be seen in his paintings. Tones in his later paintings became muddy; whites and greens became yellow [6,85]. Edgar Degas was almost blind for his last twenty years. He worked mostly in pastel with increasingly broad, free handling. Henri Matisse lost his mobility in his later years. He had to draw from his bed and let his assistant cut the paper. Like those painters, most senior citizens prefer to live in their own homes independently. Can a machine have empathy to understand human's feeling or states? What can an empathic artifact do for us at home?

For decades, computers have been viewed as apathetic machines that only accept or reject instructions. Whether an artifact can understand human's feeling or state is a paradox of empathy. René Descartes claims that thoughts, feelings, and experience are private and it is impossible for a machine to adequately understand or know the exact feelings of people. On the other hand, Ludwig Wittgenstein states that there is no way to prove that it is impossible to adequately imagine other people's feeling [44]. Alan Turing argues that machine intelligence can be tested by dialogs through a computer keyboard [53,73,79]. In our case, the Turing Test can be simplified as a *time-sharing test*, where empathic machines and humans coexist in a care-giving system with a time-sharing schedule. If a person receives care continuously, then we may call the system 'empathic'.

Empathic computing emerges as a new paradigm that enables machines to know who, what, where, when and why, so that the machines can anticipate and respond to our needs gracefully. Empathic computing in this study is narrowed down to understand the ‘low-level’ subconscious feelings, such as pain, illness, depression or anomaly. Empathic computing is a combination of Artificial Intelligence (AI), network communication and human-computer interaction (HCI) within a practical context such as healthcare.

The AI program ELIZA is perhaps the first artifact that is capable to engage in an empathic conversation [80]. Based on simple keyword matching, the program appears to be a ‘good listener’ to psychiatric patients. This shows that a small program could generate pseudo-empathy at a certain degree. However, human feelings and states are more than just verbal communication. We watch, listen, taste, smell, touch and search. Warwick’s project Cyborg [83] is probably the most daring physical empathic artifact. The pioneer implanted an electrode array under his skin that interfaced directly into the nervous system. The signal was fed into a robot arm that mimicked the dynamics of Warwick’s own arm. Furthermore, the researcher implanted a sensor array into his wife’s arm with the goal of creating a form of telepathy or empathy using Internet to communicate the signal remotely.

With the growing need for home health care, empathic computing attracts attention from many fields. Recent studies include designing a home for elderly people or people with disabilities [17]. Healthcare systems are looking for an easy and cost-effective way to collect and transmit data from a patient’s home. For example, a study [26] shows that the GSM wireless network used by most major cell phone companies was the best for sending data to hospitals from a patient’s home. Universities and corporations have launched labs to explore the healthy living environment, such as LiveNet [65,40], HomeNet [28], and Philips’ HomeLab [27]. Furthermore, Bodymedia has developed the armband wearable sensor [8,20] that tracks body temperature, galvanic skin response, heat flux, and other data. The data are then uploaded to a special web site for food and nutrition advices.

In this paper, we explore concepts of empathic sensor webs and related empathic computing methods. As a case study, we focus on the wearable sensor network for anomalous event detection at home. Using a simple distributed wireless sensor network and an equally simple algorithm, we are able to determine if the person is in trouble in real time. They may have fallen over and hurt themselves; their body temperature may be abnormal; or they may have stopped moving for an extended period of time. It would be valuable to alert the appropriate persons for assistance.

## 2 Empathic Sensor Web

Alarm pheromones are released by insects such as fish and bees when they alert others of danger [61,42,88]. Although human alarm pheromones are still debatable, there is no doubt that our instinct often makes us aware of dangerous situations. People can usually sense trouble with a car from noises, vibrations, or smells. An experienced driver can even tell where the problem is. Empathic computing aims to detect anomalous events from seemingly disconnected ambient data that we take for granted. For example, the human body is a rich ambient data source: temperature, pulses, gestures, sound, forces, moisture, et al. Also, many electronic devices provide



pervasive ambient data streams, such as mobile phones, surveillance cameras, satellite images, personal data assistants, wireless networks and so on. The peripheral vision of the redundant information enables *empathic sensor webs*.

Early sensor webs were developed for environmental monitoring [19]. For example, the geographical information system developed by Jet Proportion Laboratory provides a pervasive, continuous, embedded monitoring presence. The concept is to deploy a large number of affordable heterogeneous sensors to form a macro instrument. The growing online sensors and mobile computing enable the sensor-rich Internet for serendipitous empathic systems. Unlike traditional sensor networks, empathic sensor webs have unique characteristics: affordability and interaction. In the following sections, we will discuss these in detail.

## 2.1 Serendipitous Sensors

Mass-produced electronic devices such as cable television sets and mobile phones provide affordable platforms for *ad-hoc* sensing and communication at a low cost. This is in contrast to the traditional sensor networks, which were mainly developed in government-sponsored institutes, large corporations or the military, where the market scale is very limited. Many commercial off-the-shelf devices can be used as sensors for empathic webs. For example, a webcam can be used for monitoring the elderly when they are at home alone [48]. A mobile phone can also be a diagnostic tool. As the sounds generated by breathing in asthma patients are widely accepted as an indicator of disease activity [63,49], researchers have investigated the use of a mobile phone and electronic signal transfer by e-mail and voice mail to study tracheal breath sounds in individuals with normal lung function and patients with asthma [5]. The results suggest that mobile phone recordings clearly discriminate tracheal breath sounds in asthma patients and could be a non-invasive method of monitoring airway diseases.

For over two thousand years, physical inspection has been a unique and important diagnostic method of Traditional Chinese Medicine (TCM). Observing abnormal changes in the tongue, blood volume pulse patterns, breath smells, gestures, etc., can aid in diagnosing diseases [94]. TCM diagnosis is a black-box approach that involves only input and output data around the body. For many years, scientists have been trying to use modern technologies to unleash the ancient knowledge base. For example, the computer-based arterial blood-volume pulse analyzer is a 'rediscovery' of the diagnostic method originated from ancient TCM [24].

Visual inspection of the tongue has been a unique and important diagnostic method of Traditional Chinese Medicine (TCM) for thousands of years. Observing the abnormal changes in the tongue proper and in the tongue coating can aid in diagnosing diseases [56]. The inspection of the tongue comprises the inspection of the tongue body and the coating. The tongue body refers to the tissue of the muscle and blood vessels, while the coating refers to something on the tongue like mosses, which are formed, according to the theory of TCM, by the rising of the 'qi' (energy) of the spleen and stomach. In the study [10], the author uses a portable digital scanner to acquire the tongue image. The features on the tongue are represented as a vector of variables such as color space coordinates  $L^*a^*b$ , texture energy, entropy and fractal index, as well as crack index. With Probability Neural Network, the model reveals the correlation between the colon polyps and the features on the tongue. Although the study is preliminary, it shows the potential of inexpensive mobile cameras playing a

role in healthcare. TCM diagnosis is not a replacement of the modern diagnostic technologies such as MRI, CT, Ultrasound, DNA, but an alternative tool for early warning that brings people for further clinical diagnoses. With the growing digital technologies, it is possible to see more personal diagnostic tools in stores, just like those pregnancy test kits or diabetes self-test kits today.



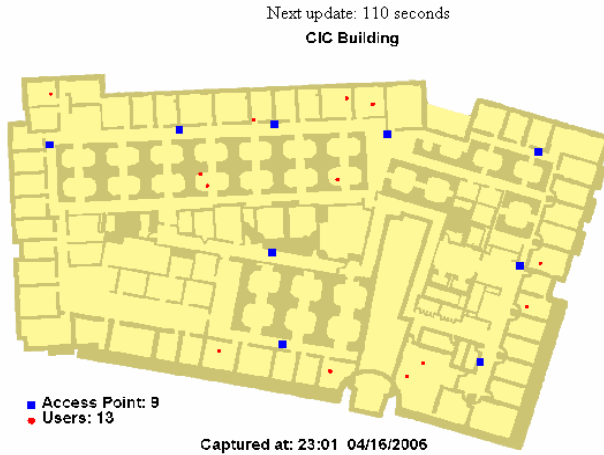
**Fig. 1.** Pocket PC-based tongue imaging system

Many existing information systems may also be used as empathic sensor webs. A wireless local network, for example, can provide a serendipitous user positioning system. Based on the Radio Signal Strength Indication (RSSI) in an indoor wireless environment, the system can estimate the distance between the access point and the wireless device. The triangulation of the user's location can be calculated from multiple access points. However, in many cases, only one access point is actively connected. Indoor furniture information and the Bayesian model are used to improve positioning accuracy with physical constraints and historical ground truth data. Figure 2 shows a screen capture of the wireless laptop positioning output. With mobile sensors such as RFID (Radio Frequency Identification) tags and multiple readers, a positioning system can also be configured in addition to the WiFi (Wireless Fidelity) system. Combined with multi-modal sensors, such as RFID, sound, infrared and magnetic signals, the positioning accuracy can be further improved.

The widely distributed open sensory system also raises serious concerns about data privacy [69]. Figure 2 shows an output from a wireless device positioning system at a building, where the location of wireless users and access points are visible on the Internet. The identity of users is replaced with a dot to preserve individual privacy.

## 2.2 Interacting with Sensor Web

For many years, sensor networks have often referred to the networked stationary sensors that are embedded inside the infrastructure systems. For example, surveillance cameras,



**Fig. 2.** Screen capture of the real-time wireless positioning system

WiFi access points, RFID readers, smart carpets, IR motion sensors, proxy sensors and so on. Stationary sensors are normally ubiquitous and fast in communication. However, they have dead zones and poor scalability. For example, for house monitoring, the more rooms in a house, the more cameras are needed. Consequently, the cost of a sensor network in a home would grow exponentially as rooms increase.

To solve the scalability problem, one solution is to enable mobile sensors interacting with stationary sensor networks. Mobile sensors include wearable or portable sensors, such as implanted RFIDs, mobile phones, e-watches, armband sensors, smart shoes or smart clothes. Mobile sensors can be configured as ad-hoc sensor networks by themselves, e.g. ZigBee ad-hoc network. These sensors are intrusive, but have a good spatial and temporal scalability if the number of people is small.



**Fig. 3.** A camera view of a nursing home with an overlay of a trajectory of the tracked human head positions (stationary sensor, left) and a prototype of the wearable infrared, temperature and motion sensors (mobile sensor, right)

### 2.3 Power Law of Resolution

A growing number and variety of sensors and other data sources are generating ever-larger volumes of data, including text, numeric, geospatial, and video data. Sensor webs would bring an ‘information avalanche’ to us that may overwhelm our data interpreting systems. The Radio Frequency Identification system is a vivid example. A large number of RFID across checkpoints in a supermarket may jam the network and computer systems. Distributed or localized data processing is necessary for sensor webs for the manageable network traffic control and computing resources. We must minimize the information flow while maximizing the perceptual capacity in a sensor web. Fortunately, the most important advantage of a sensor web is its interactivity. Multiple low-resolution sensors online may generate high valued results.

The fidelity of a sensor web can be distributed as a power law, or Pareto curve. Thus, we have about 80% low resolution and 20% high resolution. Human information processing follows the power law. If the data are plotted with the axes being logarithmic, the points would be close to a single straight line. Humans process only a very small amount of information in high fidelity, but large amounts of information in middle or low fidelity. The amount of processed information is roughly inversely proportional to its level of fidelity. Therefore, we have a fidelity power law for assigning the information processing capability for sensory channels. Given the amount of sensory information  $X$ , and the processing fidelity  $Y$ , the relation can be expressed as:

$$Y = -a \cdot \log(X) + b \quad (1)$$

where  $a$  and  $b$  are the constants. For example, we have surprisingly low visual acuity in peripheral vision (rods) but very high visual acuity in the center of gaze (cones). Curiously, despite the vitality of cones to our vision, we have 125 million rods and only 6 million cones. Our gaze vision is optimized for fine details, and our peripheral vision is optimized for coarser information.

Because network capacities are limited in comparison with those of wired networks, wireless networks are much more susceptible to overload if the wrong data is transmitted or is sent to the wrong people at the wrong time. Sending video to someone who does not want or need it not only distracts the human, but also uses up network bandwidth that cannot be used for something more useful. One approach is content routing, which attempts to move data to where it is needed for analysis or decision making without overloading wireless links. Another strategy is to anticipate the locations where many people will need to look at a particular piece of information, and then move that information to a local server for later asynchronous access.

Considering a remote sensing system for monitoring a community of elderly people, how many screens do we need for the control room? How many operators do we need for vigilance around the clock? In author’s recent study [11], eye gaze tracking and face detection technologies are applied to optimize the throughput of a wireless mobile video network. From the empirical experiments it is found that multiple resolution screen switching can reduce the network traffic about 39%. With eye gazing interface, the throughput of the network reduced about 75%. Combining the eye tracking and face detection in the video, the overall throughput reduction reaches about 88%.

### 3 Sensor-Rich Computing

Empathic sensor webs bring a new paradigm of *sensor-rich computing* that does more distributed measurements than centralized computing, which exists in nature for millions of years. For example, a modern aircraft requires about 6 million lines of code to be aware of its situations, based on a few sensors. On the other hand, a fly uses only a few hundred neurons in its brain (about 2 percent) to do the same job. About 98% of the neurons the fly uses are devoted to process near one million channels of sensory data [93]. In light of this, we could use more distributed sensors to replace the heavy-duty centralized computing. This also suggests that there is no need for high-resolution sensors. Instead, a large number of coarse-grained sensors could give reasonable results. Most biosensors are transformers that convert one kind of signals into another. A crock ranch can feel human motion by sensing the airflow that passes its hairs. Obviously, it is not capable to solve the complex fluid dynamics equations [54].

An ad-hoc sensor web may generate more data than we can handle. Most information today hasn't been analyzed even if it actually contains actionable information. Automation is essential to process, filter, and correct this flood of data, and to present it as accurate and actionable information for humans. As information from multiple sources flows up to higher levels, a more complete picture can be created, enabling adjudication at the higher level to correct erroneous information that has arisen at lower levels. Adjudication also helps reduce the volume of information being pushed up, which can overwhelm decision-makers.

## Pain Assessment Scales

Choose a number from 0 to 10 that best describes your pain



From Jacob A. et al. Rockville, MD: Agency for Health Care Policy and Research (AHCPR). US Dept. Health and Human Services. Publication No. 94-0992, 1994.

Choose the face that best describes how you feel



From Wong DL, Hockenberry-Eaton M, Wilson D, Winkelstein ML, Ahmann E, Divito-Thomas PA, Whaley and Wong's *Nursing Care of Infants and Children*, ed. 6, St. Louis, 1999, Mosby, p. 1113. Copyright Mosby. Reprinted by permission.

Fig. 4. Expressions of pain in pictures, numbers and words<sup>1</sup>

<sup>1</sup> From Hockenberry MJ, Wilson D, Winkelstein ML: *Wong's Essentials of Pediatric Nursing*, ed. 7, St. Louis, 2005, p. 1259. Used with permission. Copyright, Mosby.

### 3.1 Semantic Differential Representation

The Semantic Differential method measures perceptual and cognitive states in numbers or words. For example, the feeling of pain can be expressed with adjectives, ranging from weakest to strongest. Figure 4 shows a chart of visual, numerical and verbal expressions of pain in hospitals: No Hurt (0), Hurts Little Bit (2), Hurts Little More (4), Hurts Even More (6), Hurts Whole Lot (8) and Hurts Worst (10).

The physical feeling can be quantified with mathematical models. When the change of stimulus ( $I$ ) is very small, we won't detect the change. The minimal difference ( $\Delta I$ ) that is just noticeable is called perceptual threshold and it depends on the initial stimulus strength  $I$ . At a broad range, the normalized perceptual threshold is a constant,  $\Delta I/I = K$ . This is so-called Weber's Law [39].

Given the perceptual strength  $E$ , as the stimulus  $I$  changes  $\Delta I$ , the change of  $E$  is  $\Delta E$ . We have the relationship  $\Delta E = K * \Delta I/I$ . Let  $\Delta I$  be  $dI$  and  $\Delta E$  be  $dE$ , thus we have the so-called Weber-Fechner's Law:

$$E = K * \ln(I) + C \tag{2}$$

where,  $C$  is constant and  $K$  is Weber Ratio,  $I$  is stimulus strength and  $E$  is the perceptual strength. Weber-Fechner's Law states that the relationship between our perceptual strength and stimulus strength is a logarithm function. Studies show the values of Weber Ratios: sound strength 0.088, sound pitch 0.003, pressure 0.136, and illumination 0.016 [39].

### 3.2 Inverse Physics

Inversion is the process of retrieving physical properties [ $\mathbf{P}$ ] from observations [ $\mathbf{P} = f^{-1}(\mathbf{O})$ ]. For example, for given vibration, sound and infrared signals, the state of people in a room can be estimated from physical models [ $\mathbf{B} = f(\mathbf{W})$ ]. The simplest inversion to retrieve human state  $\mathbf{W}$  from observation  $\mathbf{B}$  [ $\mathbf{W} = f^{-1}(\mathbf{B})$ ] is a linear regression

$$W = C_1 B_{1,h} + C_2 B_{1,v} + C_3 B_{2,h} + \dots \tag{3}$$

For given physical properties, we can generate a library of high resolution simulation results for sensors. Most inverse problems are *nonlinear* in nature [74]. The generalized nonlinear regression (GNR), a machine learning model [29], can be used to estimate the human states from observations:

$$W(\vec{B}) = \sum_{i=1}^N \hat{W}_i D_i / \sum_{i=1}^N D_i \tag{4}$$

$$D_i = \exp\left[-\sum_{j=1}^M \frac{(\hat{B}_{j,i} - B_j)^2}{(\rho_j \sigma_j)^2}\right] \tag{5}$$

where  $\hat{W}_i$  and  $\hat{B}_{j,i}$  are human states and observations from forward model simulations and previous retrieval results.  $\rho$  is a correlation factor between observation channels, and  $\sigma$  is measurement error of  $\hat{B}_{j,i}$ . Based on radial-bases neural networks, GNR is a non-parametric estimation method. Retrievals using GNR are as straightforward as

linear regressions, but yield more accurate results. GNR has been tested in many remote sensing applications. With all the existing simulation models, we can produce equivalent  $\hat{W}_i$  and  $\hat{B}_i$  for interested events from forward simulation. Comparing with other inverse methods, GNR is more universal since it does not require *a priori* information. However, GNR is not necessarily an ideal candidate for embedded sensor fusion due to its large memory demand. To improve the parallelism of the algorithm, we modify the GNR to its subset, the Koheren Model [33,34], or Radial Basis Function (RBF) model, which is simpler and easier to be embedded and parallelism.

$$W(\bar{B}) = W_0 + \sum_{i=1}^k \hat{W}_i D_i \quad (6)$$

$$D_i = \exp\left[-\frac{\text{dist}(\hat{B}_i, \bar{B})^2}{2\sigma_u^2}\right] \quad (7)$$

where each  $\hat{B}_i$  is a kernel center and where  $\text{dist}()$  is a Euclidean distance calculation. The kernel function  $D_i$  is defined so that it decreases as the distance between  $B_u$  and  $B_i$  increases. Here  $k$  is a user-defined constant that specifies the number of kernel functions to be included. The Gaussian function  $D_i$  is centered at the point  $\hat{B}_i$  with some variance  $\sigma_u^2$ . The function provides a global approximation to the target function, represented by a linear combination of many local kernel functions. The value for any given kernel function is non-negligible only when the input  $\bar{B}$  falls into the region defined by its particular center and width. Thus, the network can be viewed as a smooth linear combination of many local approximations to the target function. The key advantage of RBF networks is that they contain only summation of kernel functions, rather than compounded calculation, so that RBF networks are easier to be parallelized. In addition, RBF networks can be trained with a matrix of weights so that they need less memory than GNR that stores a sequence of historical data.

### 3.3 Inversion-On-Chip

Sending raw data to a server would saturate the bandwidth of a sensor web. To solve the problem, we have also implemented the above algorithms on the Field Programmable Gate Array (FPGA), which is reconfigurable and parallel [89]. Inversion-on-chip enables us to offload the data traffic from the sensor web and synthesize the alarm pheromones in real-time. A prototype of the physical inversion models is constructed on the NI PXI-7831R FPGA prototyping board. The FPGA Vertex II 1000 contains 11,520 logic cells, 720 Kbits Block RAM, and 40 embedded 18x18 multipliers. Figure 5 shows a basic design for GNR on FPGA.

To increase the capacity and speed, we have also implemented Radial Basis Function, a subset of the Generalized Non-linear Regression model. The RBF model increased the capacity in two folders and up.

We have the following preliminary results through our benchmark experiments: 1) the FPGA chip over-performance Pentium at least two to three orders of magnitude in

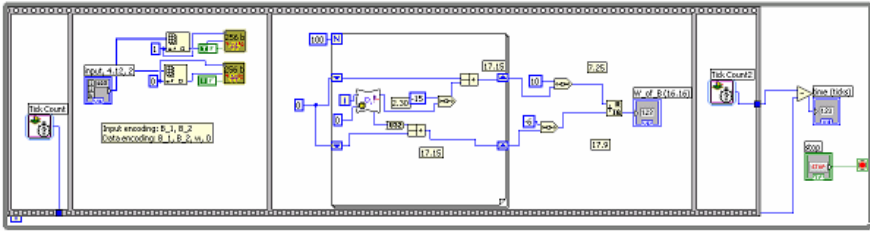


Fig. 5. Basic design for GNR implementation on FPGA

terms of speed. For example, for the GNR model, the FPGA uses 39  $\mu$ s with 10 MHz clock speed. Pentium uses between 1000  $\mu$ s and 2000  $\mu$ s with 1 GHz clock speed. 2) The fixed point Radial Basis Function algorithm demonstrated ideal parallelism as the number of simultaneous basis compares increases. Figure 6 shows the computing time for the parallel processing for GNR and RBF models (left) and the resources utilization by the two models (right).

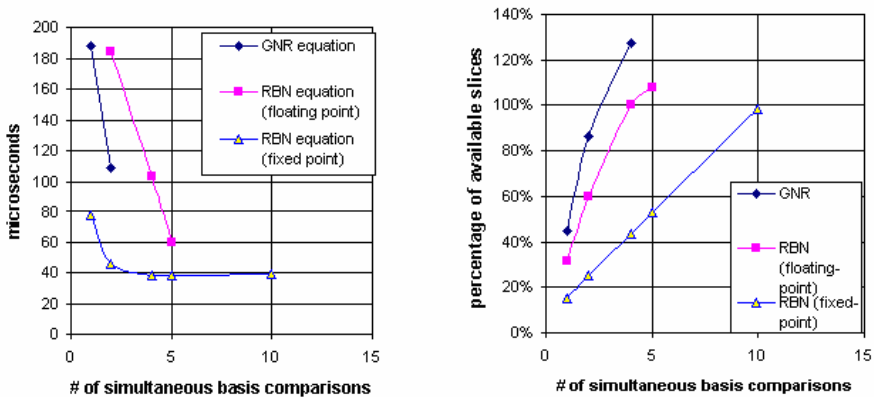


Fig. 6. Computing Time (left, 2 Inputs and 100 Basis) and Resources Utilization (right)

### 4 Anomaly Detection – Case Study

Detecting elderly people’s health states at home has been an increasing demand. There are many solutions for detecting a fall, such as passive sensor networks and active sensory networks. In our trials with wireless sensor networks, we were using a MICA2/DOT development platform [16]. These sensors are easy to set up and are able to sense temperature,  $x$  and  $y$  position, light, sound, and  $x$  and  $y$  acceleration. These nodes are compatible with TinyOS. This is an open source platform for node programming. Our initial setup is very humble: one base station connected to a computer, and one node attached to the subject (Figure 7).

This, of course, can be extended to include multiple nodes all over the body as well as all over the house. By using these tiny sensors, our goal is to determine, in the most



accurate and non-intrusive way possible, if someone has fallen down. Our algorithm must be able to process data quickly, and to distinguish a fall from other daily activities such as sitting or lying down, bending over, etc. This simple test is to show that wireless sensor networks can be used reliably to send data back to a hospital or care-giver.

#### 4.1 Sampling Rate and Sensor Placement

When choosing a sampling rate for the sensor, we need to determine how fast a person falls and also how fast the sensor could reliably send data back to the base node. In the beginning, our sensor's sampling rate was at four seconds. With this slow sampling rate, one could see that the person was standing, and then fallen, but it was hard to find where the initial fall was. We increased the sampling rate to four times per second and then we were able to see the change as the person fell. Four times per second was chosen because it gives enough data to determine a fall, but it is slow enough so that more sensors could be added and the network would not slow down.

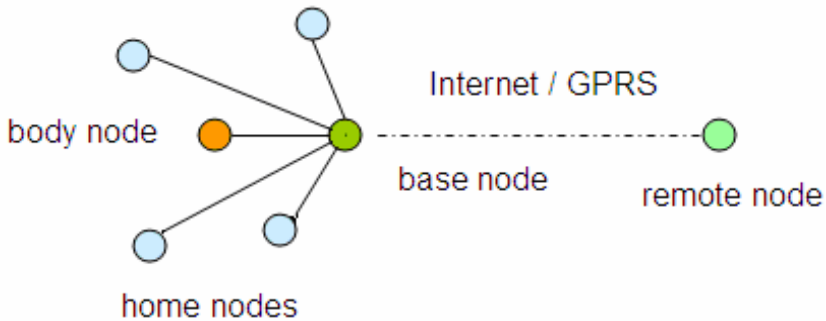
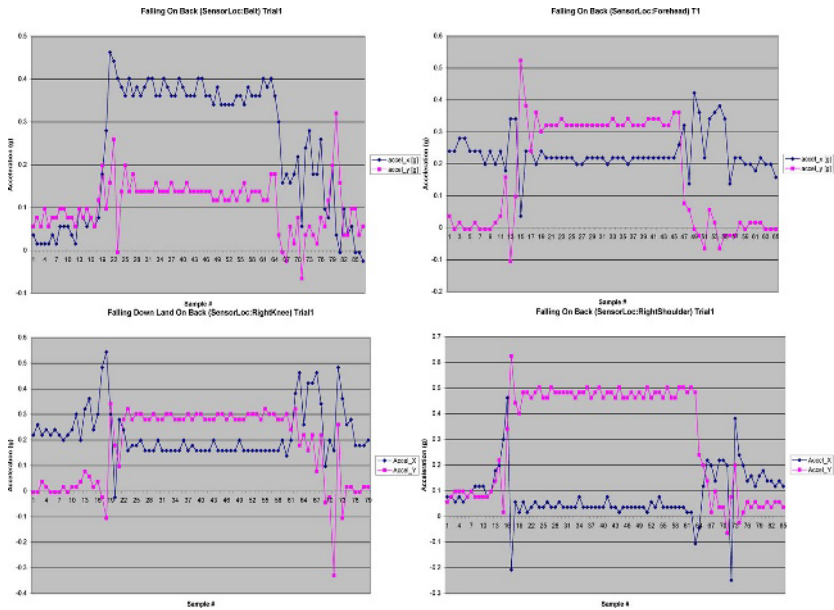


Fig. 7. The sensor web test bed



Fig. 8. Example of node placed on body

The position of the sensor was another factor that was very important. Our goal is to make the sensor invisible to the user, yet still very functional. As you will see, we tried several different locations including the knee, the belt, the shoulder, and the forehead. When the sensor was placed on different parts of the body, it would give very different readings due to the different movements of the particular body part. We wanted to pick the part of the body that gave us the cleanest readings and that was also easy for the user to wear. The belt and the forehead gave us the best readings. We decided that the belt would be more realistic for a person to wear all the time than the forehead.



**Fig. 9.** Differences in fall when sensor was placed on different parts of the body. Starting at the top left and going clockwise we have sample fall data from belt, forehead, right shoulder, right knee.

### 4.2 Empirical Studies

In our design, we set up the system to log all the data received from the nodes into a SQL compliant database. We did this for a few reasons. First of all, the data is kept in a place that can be easily accessed by programs or hospitals that need to check all data to see if there are changes over time in various areas. Another reason is that if all the data is in a database, the programs can be set up to run from anywhere in the world as long as they have Internet access. This means that the programs could be run from a hospital or a caretaker’s home and that would make reporting anomalous behavior easier. Using TinyOS, we were able to set up a program that simply reads data packets received from the nodes, parses them into relevant data, and logs them into an SQL database for us to a defined format. We had this program running all the time in order to log the data. Once we had the data, we had to determine a way to analyze it that was both fast and accurate.

The sensor was given readings at a rate of four per second. This could be increased or decreased depending on the application. We found four readings per second to be sufficient for fall detection. When we are talking about an always on, always reporting sensor, the data can get overwhelming very quickly. For this reason we must choose an algorithm that is fast and efficient as well.



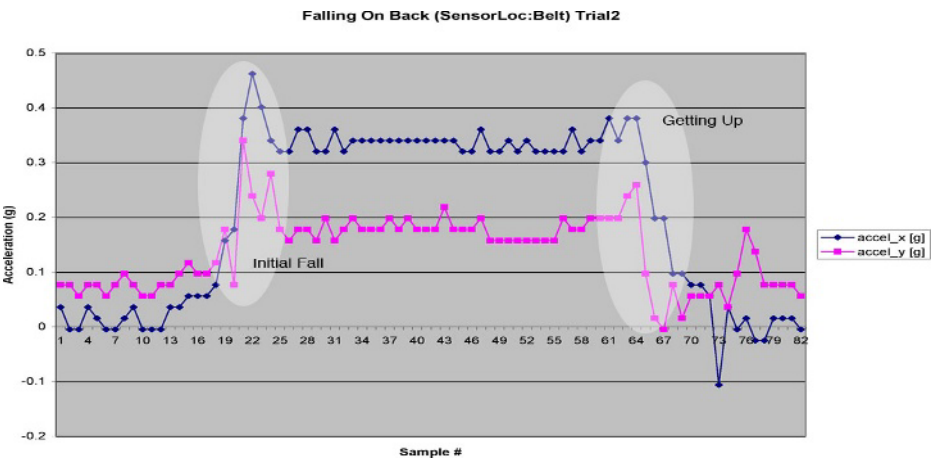
**Fig. 10.** Lab environment for simulated falls

When a person falls, the body orientation changes from vertical to horizontal. This causes the  $x$  and  $y$  acceleration to switch values and is a very good indicator that someone has fallen down. When the sensor is placed on the belt, there are very few times that a person changes  $XY$ -orientation at that speed. So we developed and trained a system of neural networks that looked for drastic changes in the  $XY$ -orientation of the subject. The neural network is a two-layer feed forward network with ten inputs and one output. Each node in the hidden layer uses the kernel function and the output uses a simple linear transfer function. The neural network was trained using the neural network stated in equation 6 and 7. This modal was used because of its relative speed and accuracy.

We then trained each modal using two walking samples, two falling samples, two have fallen samples, and two getting up samples, for a total of eight training samples for the neural network. We intend to develop a way to keep analyzing data as it comes into the database, as well as to compare it to previous data. We set up a window of ten measurements at a time. As soon as a new signal came in, we put all ten values through our series of neural networks and read the results. If the resulting value that came out of our neural networks is above a certain threshold, we would count that as a *fall*. Once a person has fallen, they cannot fall again, yet the data would continue to indicate that they had fallen until they had gotten up. This problem was easily solved with the inclusion of a flag that was set when someone fell, and reset when they got up. As long as the flag was set, they could not fall again until it was reset.

### 4.3 Detection Results

We tested our algorithm with data received from placing the node on different parts of the body. We placed the node on the forehead, the right shoulder, the belt, and the knee. Then we tested falling onto different sides of the body (chest, back, left side, right side) as well as everyday activities such as walking, sitting, and standing. We simply needed a way to determine if our method would detect everyday movement and treat it as falls. We were able to determine a fall with the most accuracy when using the data from the node placed on the belt. In second place was placing the sensor on the forehead. This is most likely because the torso or the forehead of a person stays much more stationary as opposed to the knee or the shoulder. When the sensor was placed on the belt, we were able to determine a fall almost every time. Our results are summarized in Table 1.



**Fig. 11.** Example of data received from a fall. The point where the person fell is indicated.

**Table 1.** Results of detecting the fall

| Sensor Position | Identified Correctly | Identified Incorrectly | Missed | Total Trials |
|-----------------|----------------------|------------------------|--------|--------------|
| Belt            | 90%                  | 5%                     | 5%     | 21           |
| Forehead        | 81%                  | 14%                    | 5%     | 21           |
| Right Shoulder  | 71%                  | 19%                    | 10%    | 21           |
| Right Knee      | 62%                  | 29%                    | 9%     | 21           |

The model developed in this study is not enough for detecting all the conditions. Diabetes patients, for example, often collapse slowly when the blood sugar is low. In this case, a combination of motion sensors with other sensors or interfaces would be desirable, such as a digital watch that sends a beep to the user after detecting a motionless event. The user would push a button if everything was OK. Otherwise, the wearable device will send the alarm to the network after a few inquiries.

## 5 Conclusions

Empathic computing aims to enable a computer to understand human states and feelings and to share the information across networks. Empathic sensor webs provide new opportunities to detect anomalous events and gather vital information in daily life. The widespread availability and affordability makes it easier and cheaper to link already deployed sensors such as video cameras. New sensor web capabilities can have a major impact by changing how information is used in homecare. For example, smart mirror for tongue inspection, smart room and wearable sensors for motion pattern analysis, etc.

Empathic computing brings a new paradigm to the network-centric computing, which focuses on sensor fusion and human-computer interaction. The author addresses the potential information avalanches in the empathic sensor web and proposes possible solutions for information reduction at the source side, for instance, applying the power law for multi-resolution channel design and interacting mobile sensors with stationary sensors. Ultimately, empathic computing is sensor-rich computing. In this paper, semantic differential expressions are discussed. They can be used to transform human feelings into digital forms, or vice versa. Inverse Physics methods are introduced to model the human physical states.

As a case study in this paper, the author introduces a rapid prototype of the wearable empathic computing system that is to detect the fall event. It is not a complete system. Rather, it only shows how complicated an empathic computing could be involved. From the initial results, it is found that the location of the wearable sensor makes a difference. The belt, for example, is probably the most appropriate place to put the sensor for detecting a fall. From the machine learning algorithm, the accuracy reaches up to 90% from 21 simulated trials.

The empathic sensor web concept fits into a larger picture of a smart house [12,15,17]. Wireless sensors can be placed almost anywhere and give readings about what is happening in the home. The possibilities of wireless sensor networks in the home are infinite. In addition to being unwired, wireless communications are highly and dynamically reconfigurable without physical linking, which allows the reconfiguration of communications infrastructure in real-time. Its dynamic nature makes wireless communication especially suitable for reaching areas not served well by fixed infrastructure, as well as places where the fixed infrastructure has been compromised or damaged.

Empathic computing is not limited to helping the aging society. It can be applied to solve other everyday-life problems, e.g. the empathic computing tool would remind users to take regular breaks by monitoring online duration and stress level.

The final deployment of an empathic sensor web may rely not only on technical, but also on economical and human factors. The willingness to cooperate and a willingness to make changes are critical.

## Acknowledgement

The author would like to thank the support from Russell Savage, Rafael de M. Franco, Xavier Boutonnier and Yongxiang Hu. This project is part supported by the grants from NASA ESTO-AIST Program, NASA Langley Research Center Creativity and Innovation Program, Jewish Healthcare Foundation and the Army Research Office (ARO).

## References

- [1] Aarts, E and Marzano, S. The new everyday: views on Ambient Intelligence, 010 Publishers Rotterdam; 2003
- [2] Akgul, Y.S., et at.: Automatic Extraction and Tracking of the Tongue Contours. IEEE Trans. on Medical Imaging. Vol.18, No.10, October, 1999
- [3] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: A Survey (2001)
- [4] Albowicz, Joe, Chen, Alvin, Zhang, Lixia: Recursive Position Estimation in Sensor Networks. UCLA Internet Research Laboratory.
- [5] Anderson, K., Qiu, Y., Whittaker, A.R., Lucas, Margaret: Breath sounds, asthma, and the mobile phone. Lancet. 358:1343-44. (2001)
- [6] Backhaus, W., Kliegle, R. and Werner, J. (eds): Color Vision, Walter de Gruyter, Berlin, New York, 1998
- [7] Bishop, Christopher M.: Neural Networks for Pattern Recognition. Oxford: Oxford UP, 1995.
- [8] Bodymedia: [www.bodymedia.com](http://www.bodymedia.com)
- [9] Bossert, W. H. (1963): Chemical communication among animals. Recent Progress in Hormone Research 19, 673-716.
- [10] Cai, Y et al. : Ambient Diagnostics, in Y. Cai (ed.) Ambient Intelligence for Scientific Discovery, LNAI 3345, pp. 248-262, 2005
- [11] Cai, Y. and Milcent, G.: Video-On-Demand Network Throughput Control, Proceedings of Ambient Intelligence for Everyday Life, San Sebastian, Spain, July, 2005
- [12] Chan, M., Hariton, C., Ringard, P., Campo, E.: Smart House Automation System for the Elderly and the Disabled. Systems, Man and Cybernetics, 1995. 'Intelligent Systems for the 21st Century', IEEE International Conference.
- [13] Chaney, G.R.: Do you Snore? [http://www.garnetchaney.com/help\\_for\\_snoring.shtml](http://www.garnetchaney.com/help_for_snoring.shtml)
- [14] Chen, S., Cowan, C.F.N., and Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks, Vol.2, no.2, March. (1991) 302-309
- [15] Chong, Chee-Yee, and Kumar, Srikanta P.: Sensor Networks: Evolution, Opportunity, an Challenges (2003)
- [16] Crossbow: MOTE-KIT 5x4x. 2005. Crossbow Technology Inc.
- [17] Dewsbury, Guy, Taylor, Bruce, Edge, Martin: Designing Safe Smart Home Systems for Vulnerable People.
- [18] ELIZA demo: [http://www-ai.ijs.si/cgi-bin/eliza/eliza\\_script](http://www-ai.ijs.si/cgi-bin/eliza/eliza_script)
- [19] Environmental Studies with the Sensor Web: Principles and Practice, by Kevin A. Delin, Shannon P. Jackson, David W. Johnson, Scott C. Burleigh, Richard R. Woodrow, J. Michael McAuley, James M. Dohm, Felipe Ip, Ty P.A. Ferré, Dale F. Rucker, and Victor R. Baker, Sensors, Vol. 5, 2005, pp. 103-117
- [20] Farringdon, Jonathan, and Sarah Nashold: Continuous Body Monitoring. Ambient Intelligence for Scientific Discovery (2005): 202-223.
- [21] Figueiredo, M, et al: Expanding NASA's Data Processing to Spacecraft, Computer, June, 1999
- [22] Galstyan, A.: Distributed online Localization in sensor networks using a moving target. University of Southern California. 2004
- [23] Gist, Y. J., and Lisa I. Hetzel. Department of Commerce: We are the People: Aging in the United States. N.p.: n.p., 2004.

- [24] Gunarathne, G.P. Presmasiri, Gunarathne, Tharaka R.: Arterial Blood-Volume Pulse Analyser. IEEE, Instrumentation and Measurement Technology Conference, AK, USA, May. (2002) 1249-1254
- [25] Hayes, P.J. and Ford, K.M. (1995): Turing Test Considered Harmful, Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-95), pp. 972-977, Montreal.
- [26] Herzog, A., and Lind, L.: Network Solutions for Home Health Care Applications. Linkoping University (2003)
- [27] HomeLab: <http://www.research.philips.com/technologies/misc/homelab/>
- [28] HomeNet: <http://homenet.hcii.cs.cmu.edu/>  
<http://www.coe.berkeley.edu/labnotes/1101smartbuildings.html>  
[http://www7.nationalacademies.org/cstb/wp\\_digitaldivide.pdf](http://www7.nationalacademies.org/cstb/wp_digitaldivide.pdf)
- [29] Hu, Y., B. Lin and Y. Cai: Determination of sea surface wind speed and temperature using TMI data: A General Regression Neural Network Approach. Submitted to Remote Sensing of Environment.
- [30] Jang, J.H., Kim,J.E., Park,K.M., Park,S.O., Chang,Y.S., Kim,B.Y.: Development of the Digital Tongue Inspection System with Image Analysis. Proceedings of the Second Joint EMBS/BMES Conference. Houston, TX, USA. October 23-26. (2002)
- [31] Kaiser, R.: Smart toilet a sure sign of future technology. Chicago Tribune. Saturday December 23. (2000)
- [32] Kaiser, W J., and G J. Pottie.: Wireless Integrated Network Sensors. Communication of the ACM May 2000: 51-58.
- [33] Kohonen, T.: Self-organization and Associative Memory. 2nd edition, Springer, Berlin. (1988)
- [34] Kohonen, Teuvo, Barna, Gyorgy, Chrisley, Ronald: Statistical Pattern Recognition with Neural Networks: Benchmarking Studies. Helsinki University of Technology.
- [35] Lansing, F., L. Lemmerman, A. Walton, G. Bothwell, K. Bhasin and G. Prescott, 2002: Needs for Communication and Onboard Processing in the Vision Era. 2002 IGARSS conference proceeding.
- [36] Legg, Gary. ZigBee: Wireless Technology for Low-Power Sensor Networks. 6 May 2004. TachOnline.
- [37] Lewin, M.E., and Altman, S., eds. America's Health Care Safety Net: Intact but Endangered, Institute of Medicine, 2000.
- [38] Li, Q., Rosa, Michael De, Rus, Daniela: Distributed Algorithms for Guiding Navigation across a Sensor Network. Dartmouth Department of Computer Science (2003)
- [39] Li, W.B.: Ergonomics for Interior and Furniture Design, China Forestry Press, 2001
- [40] LiveNet: <http://hd.media.mit.edu/livenet/>
- [41] Mainwaring, Alan, Polastre, Joseph, Szewczyk, Robert, Culler, David, Anderson, John: Wireless Sensor Networks for Habitat Monitoring. (2002).
- [42] McClintock, M.K. (1984). Estrous synchrony: modulation of ovarian cycle length by female pheromones. *Physiological Behavior* 32, 701-705
- [43] McDermott, M.M. et al: Functional Decline in Peripheral Arterial Disease: Associations With the Ankle Brachial Index and Leg Symptoms. *JAMA*, July. 292:453-461 (2004)
- [44] McNabb, R.: The Paradox of Empathy, *Philosophy Now*, No.52, 2005
- [45] Meribout M., Nakanishi M.; Ogura T., A parallel algorithm for real-time object recognition, *Pattern Recognition*, Sept. 2002, vol. 35, no. 9
- [46] Michell, T.: *Machine Learning*, The McGraw-Hill Companies, 1997
- [47] Moore, G.: Cramping more components onto integrated circuits. *Electronics*, Vol. 38, No. 8, April 19. (1965)
- [48] Ordonez, J.: The Web Chat That Saved Mom's Life, *Newsweek*, November 28, 2005

- [49] Pasterkamp, H., Kraman, S., S., Wodicka, G.R.: Respiratory sounds: advances beyond the stethoscope. *American Journal of Respiratory Critical Care Medicine*. 156:974-87 (1997)
- [50] Pathirana, P.N.: Node Localization using mobile robots in delay-tolerant sensor networks, July 2005
- [51] Perrig, P., R. Szewczyk, J.D., T.V. Wen, D. E. Culler, SPINS: Security Protocols for Sensor Networks, *Wireless Networks*, Volume 8, Issue 5, Sep 2002, Pages 521 – 534
- [52] Picton, P.: *Neural Networks*. 2nd edition, Palgrave, Basingstoke. (2000)
- [53] Popple, A.V.: The Turing Test as a Scientific Experiment. *Psychology*, 7(15), 1996
- [54] Rinberg, D. and Davidowitz, H. Do cockroaches ‘know’ about fluid dynamics? *Nature* 2000 Jul 27;406(6794):368
- [55] Rodgers, C.D., *Inverse Methods for Atmospheric Sounding*, World Scientific. 2002.
- [56] Schnorrenberger, C. and Schnorrenberger, B. *Pocket Atlas of Tongue Diagnosis*, Thieme, Stuttgart, New York, 2005
- [57] Sensor Web at JPL: <http://sensorwebs.jpl.nasa.gov/>
- [58] Shah, Rahul C., Rabaey, Jan M.: *Energy Aware Routing for Low Energy Ad Hoc Sensor Networks*. University of California Berkeley 2002.
- [59] Silverstrim, J.E., H. Eric W, C. Kent: Method and Apparatus for Multi-Waveform Wireless Sensor Network. Patent WO2004US33051 (5-12-2005).
- [60] Sinha, Amit, Chandrakasan, Anantha: Dynamic Power Management in Wireless Sensor Networks. *IEEE Design & Test of Computers* pg 62-74 (2001)
- [61] Smith, R.: Alarm signals in fishes. *Rev Fish Biol Fish* 2:33-63, 1992
- [62] Spath, H.: *Cluster analysis algorithms*. Ellis Horwood Ltd., Chichester. (1980)
- [63] Spiteri, M.A., Cook, D.G., Clarke, S.W.: Reliability of eliciting physical signs in examination of the chest. *Lancet*. 2:873-75. (1988)
- [64] Starner, T.: Interview with Thad Starner, by Peter Thomas. *New Scientist*. <http://www.media.mit.edu/wearables/lizzy/newsci.html> (1995)
- [65] Sung, M. and A. Pentland, MITHril LiveNet: Health and Lifestyle Networking, Workshop on Applications of Mobile Embedded Systems (WAMES'04) at Mobisys'04, Boston, MA, June, 2004
- [66] Sung, M. and Pentland, A. Minimally-Invasive Physiological Sensing for Human-Aware Interfaces, *HCI International 2005, HCII'05*
- [67] Sung, M., A. Pentland, Minimally-Invasive Physiological Sensing for Human-Aware Interfaces, *HCI International 2005, (HCII'05)*
- [68] Sung, M., C. Marci, A. Pentland: Wearable Infrastructure and Sensing for Real-time clinical Feedback Systems for Rehabilitation, *Journal of NeuroEngineering and Rehabilitation*, (JNER'05)
- [69] Tanz, O and Shaffer, J.: Wireless Local Area Network Positioning, in Y. Cai (ed.) *Ambient Intelligence for Scientific Discovery*, LNAI 3345, pp. 248-262, 2005
- [70] Taylor, C.: Simultaneous Localization, Calibration, and Tracking in an ad Hoc Sensor Network.. MIT April 2005
- [71] Tessier, R. et al: Reconfigurable Computing for digital signal processing: A survey, *The Journal of VLSI Signal Processing*, May, 2001, vol. 28, no. 1 and 2.
- [72] Tredennick, N. Get ready for reconfigurable computing, *Computer Design*, pp 55-63, 1998.
- [73] Turing, A.M.: Computing Machinery and Intelligence. *Mind*, 59: 433-460, 1950
- [74] Vann, L. and Y. Hu: A Neural Network Inversion System for Atmospheric Remote-Sensing Measurements. 2002 IEEE IMTC.
- [75] Wang, J. J.; Katz, R. B.: Plants, W. C., SRAM Based Re-Programmable FPGA for Space Applications. *IEEE trans. on nuclear science*, 1999, vol. 46, no. 6p1, pp. 1728



- [76] Wasserman, P.D.: Advanced methods in neural computing. New York, Nostrand Reinhold. (1993)
- [77] Watsuji, T., Arita, S., Shinohara, S., Kitade, T.: Medical Application of Fuzzy Theory to the Diagnostic System of Tongue Inspection in Traditional Chinese Medicine. IEEE International Fuzzy Systems Conference Proceedings. (1999) 145-148
- [78] Watt, S.N.K.: Naive Psychology and the Inverted Turing Test. *Psychology*, 7(14), 1996
- [79] Weizenbaum, J. *Computer Power and Human Reason: From Judgment To Calculation*, San Francisco: W. H. Freeman, 1976 ISBN 0-7167-0463-1
- [80] Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine, *Communications of the Association for Computing Machinery* 9 (1966): 36-45.
- [81] West, Brent W., Flikkema, Paul G., Sisk, Thomas, Koch, George W.: *Wireless Sensor Networks for Dense Spatio-temporal Monitoring of the Environment: A Case for Integrated Circuit, System, and Network Design*. Northern Arizona University.
- [82] White paper of The National Academy of Sciences. Exploring the digital divide: Charting the terrain of technology access and opportunity.
- [83] Wikipedia: Kevin Warwick, [http://en.wikipedia.org/wiki/Kevin\\_Warwick](http://en.wikipedia.org/wiki/Kevin_Warwick)
- [84] Williams, J. A., Dawood, A. S. and Visser, S. J.: FPGA-Based Cloud Detection for Real-Time Onboard Remote Sensing, IEEE ICFPT 2002, Hong Kong
- [85] Winston, R. (editor-in-chief): *Human*, DK Publishing, Inc. 2004
- [86] Wu, H., Siegel, M. and Khosla, P.: Vehicle Sound Signature Recognition by Frequency Principle Component Analysis. The Proceedings of IMTC 1998, selected in the IEEE Transaction on Instrumentation and Measurement Vol. 48, No. 5, ISSN 0018-9456. October. (1999) 1005-1009
- [87] Wu, H., Siegel, M., Stiefelhagen, R., and Yang, J.: Sensor Fusion Using Dempster-Shafer Theory. The Proceedings of IMTC 2002. Anchorage, AK, USA, May 21-23.
- [88] Wyatt, Tristram D. (2003). *Pheromones and Animal Behaviour: Communication by Smell and Taste*. Cambridge: Cambridge University Press. ISBN 0521485266.
- [89] Xilinx Products, 2003
- [90] Xu, L., et al.: Segmentation of skin cancer images. *Image and Vision Computing* 17. (1999) 65-74
- [91] Yao, P.: Comparison of TCM Tongue Images with Gastroscopy Images. Shangdong S&T Publisher, ISBN 7-5331-1849-9. in Chinese. (1996)
- [92] Yu, Zhanshou: *Feed-Forward Neural Network*. 2005.
- [93] Zbibowski, R.: Fly like a fly, *IEEE Spectrum*, November, 2005
- [94] Zhang, E.: *Diagnostics of Traditional Chinese Medicine*. Publishing House of Shanghai University of Traditional Chinese Medicine, ISBN 7-81010-125-0. in both Chinese and English. (1990)
- [95] Zhao, Jerry, Govindan, Ramesh: *Understanding Packet Delivery Performance In Dense Wireless Sensor Networks*. University of Southern California Computer Science Department (2003)
- [96] Zhou, X.S., Rui, Y., and Huang, T.S. (2003): *Exploration of Visual Data*, Kluwer Academic Publishers

# Location and Activity Recognition Using eWatch: A Wearable Sensor Platform

Uwe Maurer<sup>1</sup>, Anthony Rowe<sup>2</sup>, Asim Smailagic<sup>3</sup>, and Daniel Siewiorek<sup>3</sup>

<sup>1</sup> Computer Science Department  
Technische Universität München, Munich, Germany  
[uwe.maurer@mytum.de](mailto:uwe.maurer@mytum.de)

<sup>2</sup> Electrical and Computer Engineering Department  
Carnegie Mellon University, Pittsburgh PA, 15213, USA  
[agr@ece.cmu.edu](mailto:agr@ece.cmu.edu)

<sup>3</sup> School of Computer Science  
Carnegie Mellon University, Pittsburgh PA, 15213, USA

**Abstract.** The eWatch is a wearable sensing, notification, and computing platform built into a wrist watch form factor making it highly available, instantly viewable, ideally located for sensors, and unobtrusive to users. Bluetooth communication provides a wireless link to a cellular phone or stationary computer. eWatch senses light, motion, audio, and temperature and provides visual, audio, and tactile notification. The system provides ample processing capabilities with multiple day battery life enabling realistic user studies. This paper provides the motivation for developing a wearable computing platform, a description of the power aware hardware and software architectures, and results showing how on-line nearest neighbor classification can identify and recognize a set of frequently visited locations. We then design an activity recognition and monitoring system that identifies the user's activity in realtime using multiple sensors. We compare multiple time domain feature sets and sampling rates, and analyze the tradeoff between recognition accuracy and computational complexity. The classification accuracy on different body positions used for wearing electronic devices was evaluated.

## 1 Introduction

The eWatch is a wearable sensor and notification platform developed for context aware computing research. It fits into a wrist watch form factor making it highly available, instantly viewable, and socially acceptable. eWatch provides tactile, audio and visual notification while sensing and recording light, motion, sound and temperature. The eWatch power management was designed to operate similar to a cellular phone, requiring the user to recharge overnight. The eWatch needs to be small and energy efficient enough to allow for multiple day user studies by non-technical participants. Given these energy and size constraints, eWatch should provide the most computation and flexibility to allow an assortment of applications. The goal was to move beyond simple sensor logging and

allow for online analysis that could query the user for feedback while collecting data or provide services to showcase context aware applications.

eWatch can be used for applications such as context aware notification, elderly monitoring and fall detection, wrist PDA, or a universal interface to smart environments. The ability to sense and notify allows for a new variety of enhancements. For instance, much work has been done on fall detection for the elderly [11]. Existing systems do not function appropriately when a patient loses consciousness and cannot press a button. Current automatic systems have a high rate of false positives. An eWatch system could sense if the user was in distress and then query to confirm that it is an emergency. If the user does not respond, then the eWatch could use its networked abilities to call for help. The use of online learning could profile a patient's daily activity and notify a caretaker if a patient no longer performs their daily routines. The eWatch can also notify a patient when they should take certain medication. In order to achieve these goals, we need to accurately classify user location as well as activities.

## 2 Related Work

Several groups have developed wearable computing platforms and wearable sensor recording and processing devices [4, 2, 12]. However, most of these devices do not interact directly with the user, have insufficient battery life, or are too cumbersome for a long-term study with non-technical subjects. The idea of a smart wrist watch dates back as early as the 1930s [6] and first took a functional form with the IBM Linux Watch [10]. In its original form, the Linux Watch was a PDA on the wrist, and did not possess sensors. Later revisions of IBM's Linux Watch added acceleration and audio sensors; however, they lacked light and temperature sensors and have not targeted user context or location tracking applications. The power consumption of the Linux Watch is too great for day long operation.

Current location tracking systems offer high accuracy [14, 9] using triangulation methods; however, they require infrastructure support. In this paper we demonstrate a simple, coarse-grained location tracking method to show how eWatch can use sensor information to reason about the environment. Our method relies only on sensor samples from the environment in order to categorize the user's location.

In [1], the authors used multiple accelerometers worn on a person's body to recognize their physical activity. Sensor data from multiple body positions was combined for classifying the activities.

In [2], a low power sensor hardware system is presented, including accelerometer, light sensor, microphone, and wireless communication. Based on this hardware, a design method for a context recognition system is proposed. It evaluates multiple feature sets and makes the tradeoff between power consumption and recognition accuracy. A system that classifies household activities in realtime with a focus on low power consumption is presented in [3].

In [8], a system using an armband-based sensor array and unsupervised machine learning algorithms was able to determine a meaningful user context model.

In Section 3, we describe the sensor platform and Section 4 explains the experimental design and results of the location classification method. Section 5 describes the activity recognition method and presents the results of the data analysis. Section 5.4 addresses the performance of our on-board activity classifier.

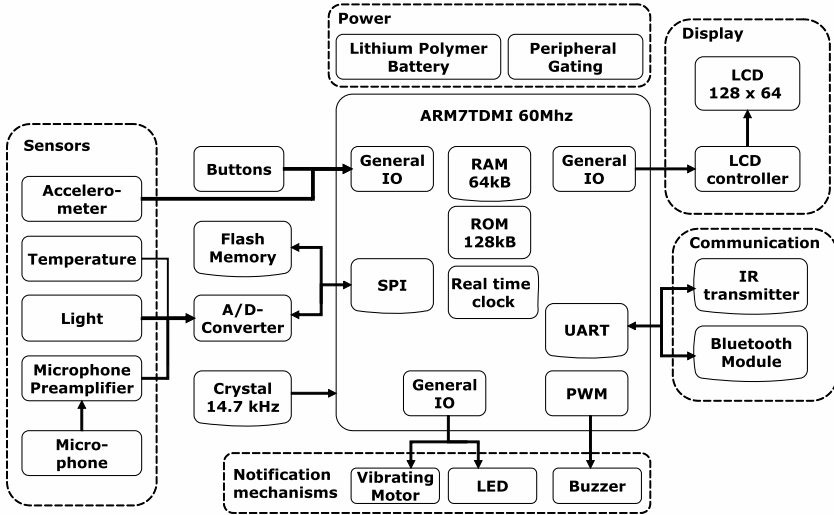


Fig. 1. eWatch hardware architecture

### 3 eWatch Design

In the following sections we will discuss the hardware and software architecture of the eWatch.

#### 3.1 eWatch Hardware

Figure 1 shows the eWatch architecture which consists of: the main CPU, sensors, power control, notification mechanisms, and wireless communication. The main CPU is a Philips LPC2106 ARM7TDMI microcontroller with 128Kb of internal FLASH and 64Kb of RAM. The LPC2106 is a 32bit processor capable of software controlled CPU scaling up to 60Mhz. eWatch communicates wirelessly using a SMARTM Bluetooth module and an infrared data port for control of devices such as a television. A previous version of eWatch based on different hardware, is described in [13].

Sensor data is acquired using an external TLV1544 10bit ADC and can be stored in a 1Mb external FLASH device. eWatch is capable of sensing temperature, light, two axes of acceleration and audio at user controllable sampling intervals up to 100Khz. A MAX4061 amplifier is used for audio conditioning.

We use an ADXL202 MEMS accelerometer to measure the planar acceleration of the user's hand. The user can be notified using a 128x64 pixel display, an LED, vibrating motor and tone generating buzzer. Three push buttons are distributed around the outside of the housing in the standard configuration of a digital watch.

eWatch is powered by a 3.6 volt 700mAh rechargeable lithium polymer battery with a linear regulator active during peak voltages and a DC to DC voltage pump as the battery drains. Table 1 shows the power consumption of each component in the system. The chart shows the maximum possible power usage of each device, followed by an average power consumption computed from a trace of daily use.

**Table 1.** Average and peak power

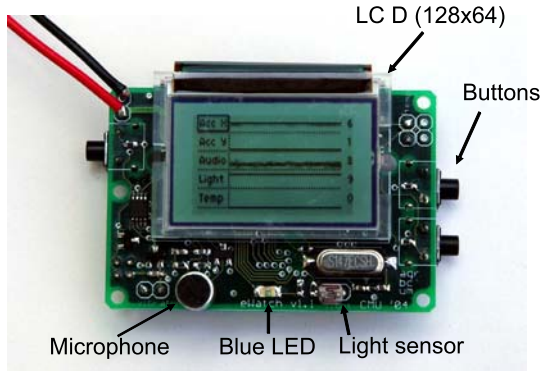
| Part                        | Avg. Power(mW) | Peak Power(mW) |
|-----------------------------|----------------|----------------|
| ARM7 processor              | 29.7           | 132            |
| ADC                         | 4.95           | 4.95           |
| Microphone Amp              | 2.5            | 2.5            |
| Accelerometer               | 2.0            | 2.0            |
| LCD Controller              | 0.23           | 0.33           |
| Serial Flash Memory         | 0.003          | 13.2           |
| Bluetooth Module            | 0.09           | 90             |
| Vibration Motor             | 0              | 63             |
| Backlight LED               | 0.03           | 33             |
| Light Sensor                | 0.825          | 0.825          |
| Temperature Sensor          | 0.825          | 0.825          |
| Average Life Time: 56 hours |                |                |

The final housing is made from epoxy resin that was cast in a silicon mold and measures 50mm x 48mm x 17.5mm. The limiting factor with respect to eWatch's size is the battery, which can later be reduced as the device is customized for a specific application.

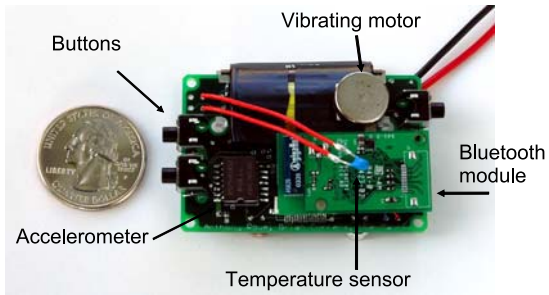
### 3.2 eWatch Software

The eWatch system was designed as a platform for developing context aware applications. The main goals that influenced the design decisions were *ease of use* and *flexibility*. eWatch provides the developer with an API that enables rapid prototyping. The eWatch software system consists of three layers: Application, System Functionality, and Hardware Abstraction.

Applications access functionality of lower layers to render screen images, interact with the user and retrieve information from the storage, sensors or wireless network. The System Functionality Layer provides an API for shell, task and power management. The Hardware Abstraction Layer contains the drivers for all the hardware components providing access to all eWatch functionality.



(a) Top view of eWatch board



(b) Bottom view of eWatch board



(c) eWatch with housing

**Fig. 2.** eWatch board and housing

The layered architecture helps achieve our goal of flexibility by reducing the effort necessary to port to another hardware or software environment. For example, we developed a Linux port of the software system that replaces the hardware abstraction layer with simulated hardware. This enables rapid development cycles, since the code can be tested on the developers' machine without actually updating the eWatch firmware.

### 3.3 Interface

eWatch offers two interfaces for a user or developer to control its functionality: the eWatch shell and the Graphical User Interface (GUI) on the built-in display.

The eWatch shell allows users to execute functions and configure variables via Bluetooth. A text-based protocol is used to transmit commands similar to a Unix shell. The applications on eWatch can register functions and variables, making them accessible through the shell. The commands can be typed by a user or developer through a keyboard or sent from a program running on the PC. This enables automated scripting of the system and allows remote applications to access eWatch functionality.

The primary GUI of eWatch is the menu system. As shown in Figure 3(a), the menus allow the user to scroll through lists of items and select entries to activate them. The menu structure is organized hierarchically - entries can be modified, added, or removed during runtime. Each menu entry is linked to a shell command that is executed when the entry is selected. The eWatch GUI library supports TrueType fonts, drawing of geometric shapes, and displaying bitmaps.

### 3.4 Applications

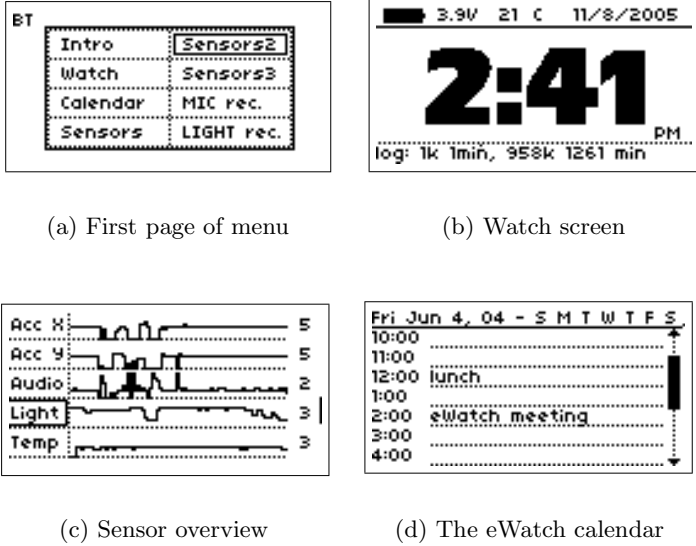
Several applications make use eWatch's capabilities. Figure 3 shows screenshots of the eWatch user interface and applications.

Like a standard wrist watch, eWatch provides the user with the current time (Figure 3(b)). In addition, it shows information about incoming text messages and upcoming calendar events. Information about the sensor logging and available memory is shown while logging is in progress. Figure 3(c) shows a real-time plot of sensor data.

The calendar application stores and notifies the user of events (Figure 3(d)). The calendar can be synchronized with a personal computer using the standard iCal format. The data from the calendar can be used for context aware applications that recognize the user's location and activity.

### 3.5 Sensor Sampling and Recording

Recording sensor data is a core system functionality. The sensor sampling and recording system is designed to consume minimal power and make efficient use of memory. Sensor sampling is interrupt-based to minimize sampling jitter. Every sensor has a timer interrupt with a configurable sampling frequency. In the interrupt handler, the ADC value is read and then stored in a memory buffer.



**Fig. 3.** The eWatch GUI

Between interrupts the system remains primarily in the idle state to conserve energy, occasionally powering up to compress the collected data and write it to flash.

We wanted a lossless compressor to allow for 24 hours of data recording. We chose an algorithm described in [7] that performs compression using four linear predictors and efficient coding of residuals. Our experiments showed that it reduced the memory consumption of the sensor data to 20% - 50% of the original size.

### 3.6 Power Management

The ARM7TDMI microcontroller supports two power-saving modes and frequency scaling. An event-based architecture that waits in idle mode for incoming events was selected. When an event occurs, the processor wakes up to service it. After the application completes, it relinquishes control to the scheduler that can then return the processor to idle mode.

## 4 Sensor Based Location Recognition

Knowing about the user's location is an important aspect of a context aware system. Using eWatch we developed a system that identifies previously visited locations. Our method uses information from the audio and light sensor to learn and distinguish different environments.

We recorded and analyzed the audio environment and the light conditions at several different locations. Experiments showed that locations have unique back-



ground noises such as car traffic, talking, noise of computers, air conditioning and television. The light sensor sampled at a high frequency can also provide additional information beyond the brightness of the location. Frequency components generated by electric lights (at 120Hz and 240Hz) and displays (television, computer screen at 60Hz) can be observed. We observed that the frequency characteristics of light conditions tend to remain constant in most locations. For our study, audio data was recorded with the built-in microphone at a sample rate of 8kHz and the light sensor at a frequency of 2048Hz. At every location, five consecutive recordings of audio and light were taken, separated by 10 second pauses. For every recording, we sampled the microphone for four seconds (32000 samples) and the light sensor for 0.5 seconds (1024 samples).

The recorded data was then compressed and stored into flash memory. Locations frequently visited by the user were recorded; the rooms of the user's apartment (living room, kitchen, bedroom, bathroom), their office, the lab, different street locations on the way to the university, the interior of a bus, and several restaurants and supermarket. Each location was visited multiple times on different days. In total, we collected 600 recordings at 18 different locations.

#### 4.1 Location Feature Extraction

We estimated the power spectral density of the recorded sensor data using Welch's method. A 128-point FFT was calculated for a sliding window over the complete recording and averaged over frequency domain coefficients for all windows. The result is a smoothed estimation of the power spectral density. To reduce the number of feature components, the Principal Component Analysis was used. The dimensionality of the feature vector was reduced to its first five principal components. To visualize the feature space, Figure 4 shows the first three components of the feature vectors after a Linear Discriminant Analysis (LDA) transformation.

#### 4.2 Location Recognition Results

The nearest neighbor method with a 5-fold cross validation was used for classification. Three different feature sets were evaluated: features from the light sensor only, microphone only and both sensors combined. As expected, the combination of both sensors gave the best results in identifying the location. The classification with the light sensor alone gave an overall result of 84.9% correctly classified samples. The classifier confused the *lab* and *office* location and also the *bus* with the *street*. This occurred because both location pairs can have similar light conditions. Using only the audio sensor, the overall recognition accuracy was 87.4%. The *office* and *apartment* location were confused in this case. Both sensors combined gave the best result of 91.4%. Locations that could not be distinguished well with only one sensor were classified more accurately with both sensors combined. Table 2 shows an overview of the classification results for the individual locations.

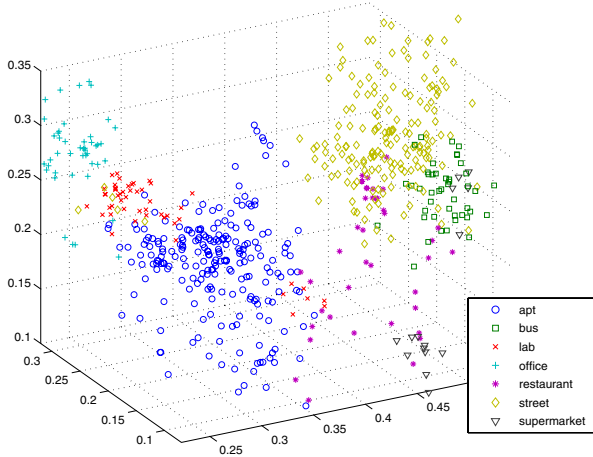


Fig. 4. Location Feature Space after Linear Discriminant Analysis Transformation

Table 2. Location recognition accuracy

| Location      | Sensors    |            |               |
|---------------|------------|------------|---------------|
|               | Light only | Audio only | Audio & Light |
| apartment     | 98.2%      | 90.9%      | 95.9%         |
| bus           | 35.6%      | 84.4%      | 77.8%         |
| lab           | 64.0%      | 90.0%      | 98.0%         |
| office        | 84.6%      | 76.9%      | 89.2%         |
| restaurant    | 62.5%      | 77.5%      | 90.0%         |
| street        | 93.5%      | 91.2%      | 90.6%         |
| supermarket   | 73.3%      | 66.7%      | 66.7%         |
| Class average | 73.1%      | 82.5%      | 86.9%         |
| Overall       | 84.9%      | 87.4%      | 91.4%         |

### 4.3 Online Classification and Performance

The 1-NN classification method was implemented on the eWatch to allow online classification in realtime. The sensor recording uses 4.5 seconds of data (4 seconds for audio, 0.5 seconds for light), the computing time for the classification is about 1.4 seconds. 98.5% of the classification time is spent performing the feature extraction. The PCA and nearest neighbor search take less than 20ms to compute. In order to reduce the time spent in the feature extraction, other features are being investigated, such as time domain characteristics.

## 5 Activity Recognition on Different Body Positions

Previous feature extraction studies examining accelerometer data have shown that it is a viable input for detecting user states when it is worn on the wrist [1].

Motivated by other possible sensor platform locations, especially with mobile communication devices such as a cell phone or PDA, we designed a study to investigate the dependency of the eWatch classification accuracy on different given body positions. We investigate wearing the eWatch in the following locations: the belt, shirt pocket, trouser pocket, backpack, and necklace. The results of the study would help us decide on the best position to place such a sensor platform, and understand the nature of the trade-off between wearing position and classification performance.

### 5.1 Activity Recognition Experiment

In our study we focussed on six primary activities: *sitting*, *standing*, *walking*, *ascending stairs*, *descending stairs* and *running*. Body positions that are normally used for wearing electronic devices, such as cell phones or PDAs, were studied. We placed our sensor hardware on the left wrist, belt, necklace, in the right trouser pocket, shirt pocket, and bag. The subjects wore six eWatch devices located at these body positions during the study. The devices recorded sensor data from the accelerometer and light sensor into their flash memory. The user was asked to perform tasks that consist of activities such as working on the computer or walking to another building. The lead experimenter annotated the current activity and instructed the subjects on how to proceed. The annotations were done using an application running on an extra eWatch worn by the lead experimenter.

Six subjects participated in the study; each subject performed the given tasks in 45 to 50 minutes. In total, we collected over 290 minutes of sensor data.

*Sensor Setup.* eWatch recorded both axes of the accelerometer and the light sensor. All sensors values were recorded with a frequency of 50Hz and with 8bit resolution. The accelerometer was calibrated so that both axes operate in a range of  $\pm 2g$ . Evaluation of the recorded data was done with Matlab and the WEKA software [15].

### 5.2 Activity Feature Extraction

The sensor values recorded from the accelerometers and the light sensor are split into short time windows. These windows are then transformed into the feature space by calculating several feature functions over the individual windows.

*Features.* Features from both accelerometer axes ( $X$  &  $Y$ ), the light sensor, and a combined value of both accelerometer signals were calculated. To reduce the dependency on the orientation, both  $X$  and  $Y$  values were combined calculating the squared length of the acceleration vector. The classification accuracy with individual sensors, as well as with multiple combined sensors, was investigated.

Only time domain features were considered to avoid the costly computation that is required to transform the signal into the frequency domain. Table 3 shows the list of features that were considered. The functions to calculate these features were implemented on the eWatch and the required number of clock cycles per

function was measured. Each function was executed 2000 times with different recorded sensor inputs, and then the average value was computed. The execution time was calculated based on the measured clock cycles and the CPU frequency at 59MHz. Table 3 shows the measured clock cycles and execution time using a four second window sampled at 20Hz (80 samples).

**Table 3.** List of time domain features and the average clock cycles and time to calculate them on the eWatch running at 59MHz

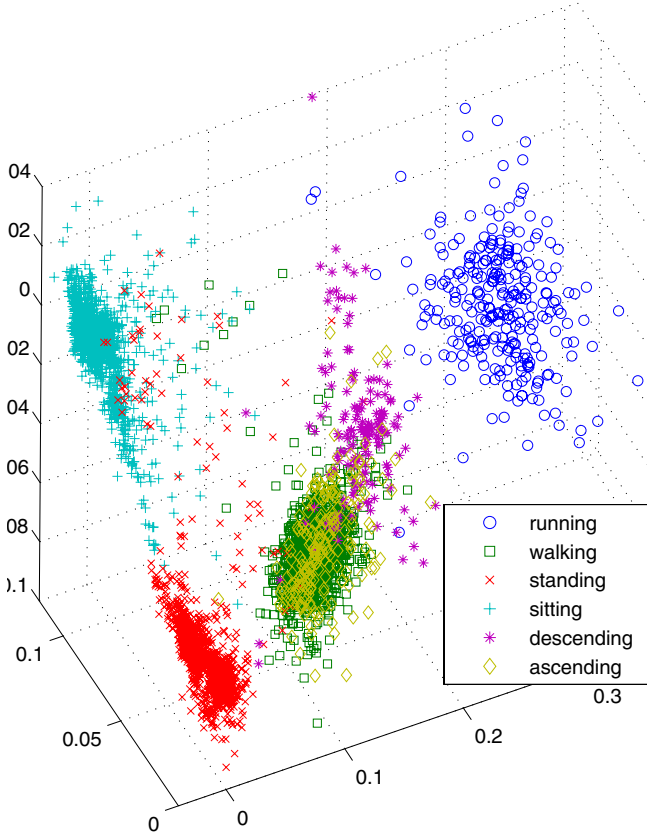
| Features / Function                        | Name        | Avg. CPU Cycles | Avg. Time in $\mu s$ |
|--|-------------|-----------------|----------------------|
| Empirical Mean                             | <i>mean</i> | 854             | 14.5                 |
| Root Mean Square                           | <i>rms</i>  | 1219            | 20.7                 |
| Standard Deviation                         | <i>std</i>  | 1139            | 19.3                 |
| Variance                                   | <i>var</i>  | 1313            | 22.3                 |
| Mean Absolute Deviation                    | <i>mad</i>  | 1089            | 18.5                 |
| Cumulative Histogram (256 bins)            | <i>hist</i> | 5847            | 99.1                 |
| n'th Percentile ( $n = 5, 10, \dots, 95$ ) | <i>prc</i>  | 142             | 2.4                  |
| Interquartile Range                        | <i>iqr</i>  | 289             | 4.9                  |
| Zero Crossing Rate                         | <i>zcr</i>  | 993             | 16.8                 |
| Mean Crossing Rate                         | <i>mcr</i>  | 996             | 16.9                 |
| Sq. Length of X,Y ( $x^2 + y^2$ )          |             | 1318            | 22.3                 |
| Decision Tree classifier (18 nodes)        |             | 138             | 2.3                  |

Figure 5 depicts the feature space after a transformation with Linear Discriminant Analysis (LDA). It shows that the *standing*, *sitting* and *running* activities form separate clusters, while *walking*, *ascending* and *descending stairs* are closer together since these activities are very similar.

*Feature Subsets.* To reduce the time and energy required to calculate the feature vector, several subsets of the complete feature space were evaluated. Some features are irrelevant or redundant and do not provide information to significantly improve the classification accuracy. Therefore, a subset of the available features can be selected to decrease the computation time without significantly decreasing recognition accuracy.

The Correlation based Feature Selection (CFS) method from the WEKA toolkit was used to find feature sets containing features that are highly correlated within the particular class, but are uncorrelated with each other. Table 4 shows the feature sets that were compared.

*Classification Method.* We evaluated and compared several classification methods, namely Decision Trees (C4.5 algorithm), k-Nearest Neighbor (k-NN), Naive-Bayes and the Bayes Net classifier. Decision Trees and Naive-Bayes were found to achieve high recognition accuracy with acceptable computational complexity. Decision Trees were used for activity classification in [1] and [2]. It was

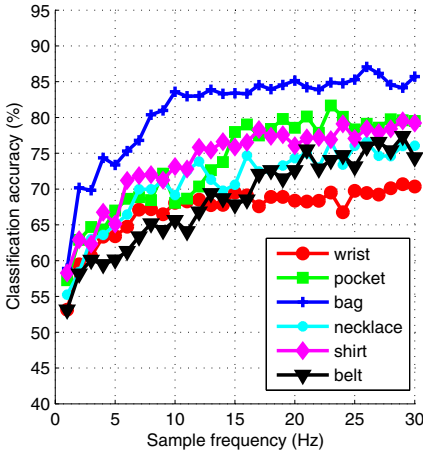


**Fig. 5.** Activity Feature Space after Linear Discriminant Analysis Transformation

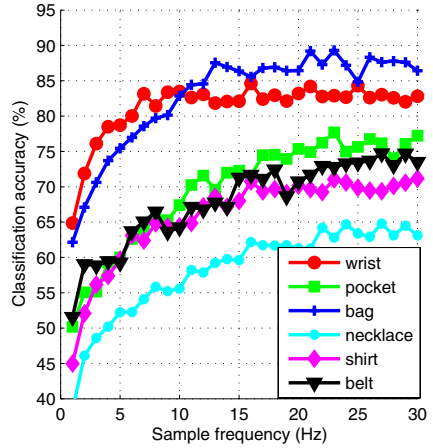
**Table 4.** Feature sub sets and processing time

| #  | Features   | CPU Cycles | Time in $\mu s$ |
|----|--|------------|-----------------|
| F1 | All features, all sensors  | 56242      | 953.6           |
| F2 | All features from accelerometer X  | 14731      | 249.8           |
| F3 | All features from accelerometer Y  | 14731      | 249.8           |
| F4 | All features from light  | 14731      | 249.8           |
| F5 | All features from accelerometer XY ( $x^2 + y^2$ )                                     | 15049      | 255.2           |
| F6 | $prc_y(3), rms_{xy}, prc_y(20), prc_y(97),$<br>$rms_{light}, mad_x, mean_y, prc_y(10)$ | 12114      | 205.4           |
| F8 | $prc_y(3), iqr_y, prc_y(10), prc_y(97), mad_x$   | 7651       | 129.7           |
| F9 | $rms_{xy}, qrt_x, rms_x, mad_{xy}, mean_{xy}$  | 10746      | 182.2           |

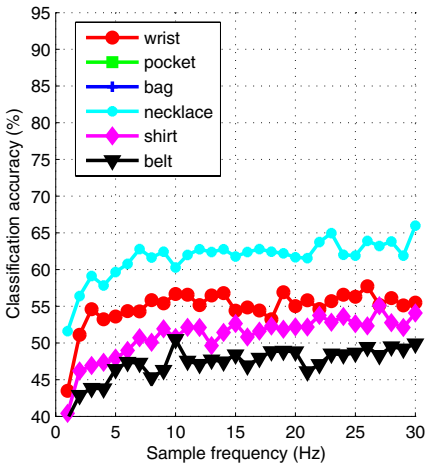
shown in [5] that the discretized version of Naive-Bayes can outperform the Decision Tree classifier for general classification problems. Finally the Decision Tree classifier was chosen as it provides a good balance between accuracy and



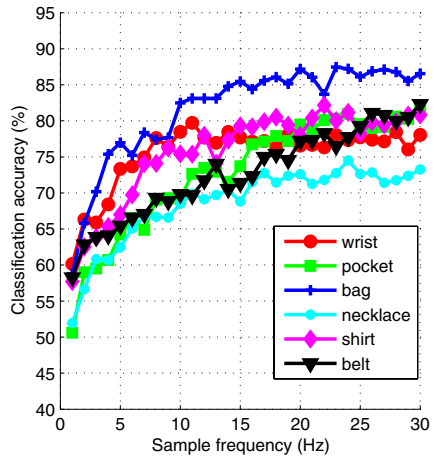
(a) Features from the accelerometer's X-axis



(b) Features from the accelerometer's Y-Axis



(c) Features from the light sensor



(d) Features from the  $x^2 + y^2$  value of the accelerometers

**Fig. 6.** Recognition accuracy with different feature sets

computational complexity. For all further experiments, this classifier with a 5-fold cross validation was used.

*Sampling Frequency.* During the user study, the sensors were sampled with a frequency of 50Hz and later downsampled to lower frequencies. To maintain some of the high frequency components' information and to reduce the computational

complexity significantly, no low pass filter was used for downsampling the data. Figure 6 shows the recognition accuracy for different sample rates from 1 to 30Hz for the different body positions. The recognition accuracy was defined as the percentage of correctly classified feature vectors averaged for all six activities. The recognition accuracy increases with higher sampling rates, and with the accelerometer features the accuracy then stabilizes between 15 to 20Hz, and is only improved marginally with higher sampling rates. The accuracy with the light sensor only is lower and it stabilizes beginning with 7Hz. In Figure 6(c) the results from the belt and pocket position are not shown because the light sensor did not provide any useful classification information at these positions.

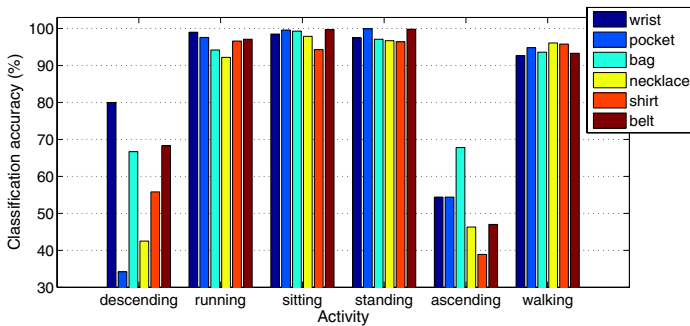
### 5.3 Activity Recognition Results

We calculated the classification accuracy for every activity on each of the six different body positions. The data from all subjects was combined to train a general classifier that is not specific to a person.

Table 5 shows the feature sets and the classification results for the different body positions. The features are calculated from a 20Hz signal.

**Table 5.** Classification accuracy for body positions

| Features | Classification Accuracy for Body Position |        |       |          |       |       |
|----------|---|--------|-------|----------|-------|-------|
|          | wrist                                     | pocket | bag   | necklace | shirt | belt  |
| F1       | 87.1%                                     | 85.2%  | 92.8% | 86.8%    | 89.5% | 87.0% |
| F2       | 68.4%                                     | 78.6%  | 85.2% | 74.3%    | 76.1% | 72.6% |
| F3       | 83.2%                                     | 75.4%  | 86.4% | 61.3%    | 70.1% | 70.8% |
| F4       | 55.0%                                     | 16.7%  | 18.0% | 61.7%    | 52.2% | 48.8% |
| F5       | 76.6%                                     | 79.5%  | 87.2% | 72.6%    | 78.0% | 77.2% |
| F6       | 87.0%                                     | 80.1%  | 86.5% | 78.6%    | 79.6% | 84.2% |
| F8       | 82.0%                                     | 62.4%  | 68.9% | 56.6%    | 69.8% | 71.7% |
| F9       | 77.3%                                     | 78.2%  | 80.9% | 72.3%    | 75.4% | 76.5% |



**Fig. 7.** Recognition accuracy for the activities at different body locations

Figure 7 shows the recognition accuracy for the individual activities at different body locations. For the classification, the reduced feature set F6 was used. The data indicate that any of the six positions are good for detecting *walking*, *standing*, *sitting* and *running*. *Ascending* and *descending* the stairs is difficult to distinguish from *walking* in all positions, since the classifier was trained for multiple persons. The wrist performs best because the feature set was optimized for the wrist position.

### 5.4 Onboard Activity Classifier

Based on these results, we implemented a decision tree classifier that runs on the eWatch. The feature set F6 was used to build the decision tree. The sensor sampling is interrupt-based, and triggers the sampling of the sensors at 20Hz. The sensor value is stored in a buffer with the size of the sliding window. The activity is classified every 0.5 seconds based on the sensor data from the 4 second buffer. The classification results are stored into flash memory and are downloaded to a computer later for further processing and analysis. They can also be transferred in realtime over the Bluetooth connection. In order to save energy, the system remains idle between servicing interrupts.

A subject wore the eWatch with the built-in activity classifier on the wrist during the day. The system classified the activity in realtime and recorded the classification results to flash memory. Figure 8 shows 100 minutes of activity classification, as the user walked to a restaurant, sat down, ate lunch, went back to the office and sat down to continue working. The classification results match well with the actual activities; eating lunch was partially interpreted as *walking* or *running* activity due to arm movements.

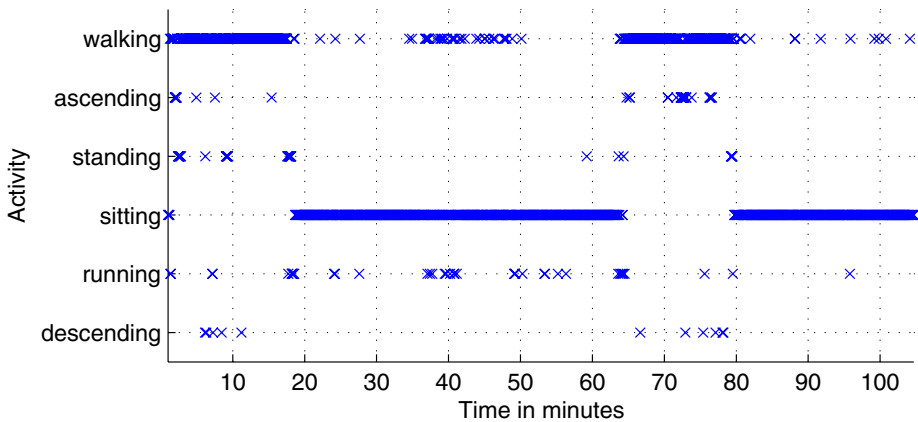


Fig. 8. Activity classification recorded over 100 minutes



## 6 Conclusions and Future Work

This paper describes eWatch, a wearable sensor and notification computing platform for context aware research. The hardware design is focused on providing enough computational resources to perform machine learning algorithms locally, while still allowing a comfortable form factor and a battery capacity sufficient for extended user studies. Likewise, the software environment was designed to facilitate easy development while automatically managing resources such as power and sensor data. We also described a system that uses the eWatch and its sensors to categorize its environment in real-time.

The activity recognition and monitoring system that can identify and record the user's activity in realtime, using multiple sensors, is presented. We compared multiple feature sets and sampling rates to find an optimized classification method, and showed how well they perform on different body locations that are commonly used for wearing electronic devices.

We will extend our activity classifier to other activities and investigate how the activity classification can support the recognition of the user's location. Integration of an 802.15.4 radio is planned to allow the eWatch to function as a mobile node in a sensor network. This added flexibility will further integrate the eWatch into its environment by allowing a larger area of network coverage.

## Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010, the National Science Foundation under Grant Nos. 0205266 and 0203448, a grant from Intel Corporation, and PA Infrastructure (PITA).

## References

- [1] BAO, L., AND INTILLE, S. S. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive* (2004), A. Ferscha and F. Mattern, Eds., vol. 3001 of *Lecture Notes in Computer Science*, Springer, pp. 1–17.
- [2] BHARATULA, N. B., STÄGER, M., LUKOWICZ, P., AND TRÖSTER, G. Empirical Study of Design Choices in Multi-Sensor Context Recognition Systems. In *IFAWC: 2nd International Forum on Applied Wearable Computing* (Mar. 2005), pp. 79–93.
- [3] BHARATULA, N. B., STÄGER, M., LUKOWICZ, P., AND TRÖSTER, G. Power and Size Optimized Multisensor Context Recognition Platform. In *ISWC 2005: Proceedings of the 9th IEEE International Symposium on Wearable Computers* (Oct. 2005), pp. 194–195.
- [4] DEVAUL, R. W., AND PENTLAND, S. The MITHril Real-Time Context Engine and Activity Classification. Tech. rep., MIT Media Lab, 2003.
- [5] DOUGHERTY, J., KOHAVI, R., AND SAHAMI, M. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning* (1995), pp. 194–202.

- [6] GOULD, C. Dick Tracy. *Comic, Chicago Tribune* (Oct 1931).
- [7] HANS, M., AND SCHAFER, R. Lossless Compression of Digital Audio. *IEEE Signal Processing 18* (July 2001), 21–32.
- [8] KRAUSE, A., SIEWIOREK, D. P., SMAILAGIC, A., AND FARRINGTON, J. Unsupervised, dynamic identification of physiological and activity context in wearable computing. In *Seventh IEEE International Symposium on Wearable Computers (ISWC'03)* (2003), pp. 88–97.
- [9] LORINCZ, K., AND WELSH, M. Motetrack: A robust, decentralized approach to rf-based location tracking. In *Proceedings of the International Workshop on Location and Context-Awareness (LoCA 2005) at Pervasive 2005* (May 2005).
- [10] NARAYANASWAMI, C., AND RAGHUNATH, M. T. Application design for a smart watch with a high resolution display. In *ISWC '00: Proceedings of the 4th IEEE International Symposium on Wearable Computers* (Washington, DC, USA, 2000), p. 7.
- [11] NOURY, N. A smart sensor for the remote follow up of activity and fall detection of the elderly. In *Proc. of the IEEE Special Topic Conference on Microtechnologies in Medicine & Biology May* (2002).
- [12] SMAILAGIC, A., AND SIEWIOREK, D. P. Wearable and Context Aware Computers: Application Design. In *IEEE Pervasive Computing, Vol. 1, No. 4* (Dec 2002), pp. 20–29.
- [13] SMAILAGIC, A., SIEWIOREK, D. P., MAURER, U., ROWE, A., AND TANG, K. eWatch: Context-Sensitive Design Case Study. In *In Proc. of the IEEE Annual VLSI Symposium* (May 2005), IEEE Computer Society Press, pp. 98–103.
- [14] WANT, R., HOPPER, A., FALCAO, V., AND GIBBONS, J. The active badge location system. Tech. Rep. 92.1, ORL, 24a Trumpington Street, Cambridge CB2 1QA, 1992.
- [15] WITTEN, I. H., AND FRANK, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

# Co-Creation in Ambient Narratives

Mark van Doorn<sup>1</sup> and Arjen P. de Vries<sup>2</sup>

<sup>1</sup> Philips Research,

Prof. Holstlaan 4,

5656 AA Eindhoven, The Netherlands

`mark.van.doorn@philips.com`

<sup>2</sup> Centre for Mathematics and Computer Science,

Kruislaan 413,

1098 SJ Amsterdam, The Netherlands

`arjen.de.vries@cwi.nl`

## 1 Introduction

Ambient Intelligence [1] aims to improve the quality of people's life by making everyday activities more convenient and enjoyable with digital media. Technically, Ambient Intelligence refers to the presence of a digital environment that is sensitive, adaptive, and responsive to the presence of people. Electronic devices are embedded in furniture, clothing or other parts of the environment; the technology recedes into the background of our everyday lives until only the function (i.e., the user interface) remains visible to people. At the same time, the human moves into the center of attention, in control of the devices around him. These work in concert to support the performance of everyday activities in an intelligent manner.

Producing such smart ubiquitous computing environments on a large scale is problematic however. First, it is technologically not possible in the near foreseeable future to mass produce a product or service that generates Ambient Intelligence, given the current state-of-the-art in machine learning and artificial intelligence. Second, it is economically not feasible to manually design and produce Ambient Intelligence applications for each person individually.

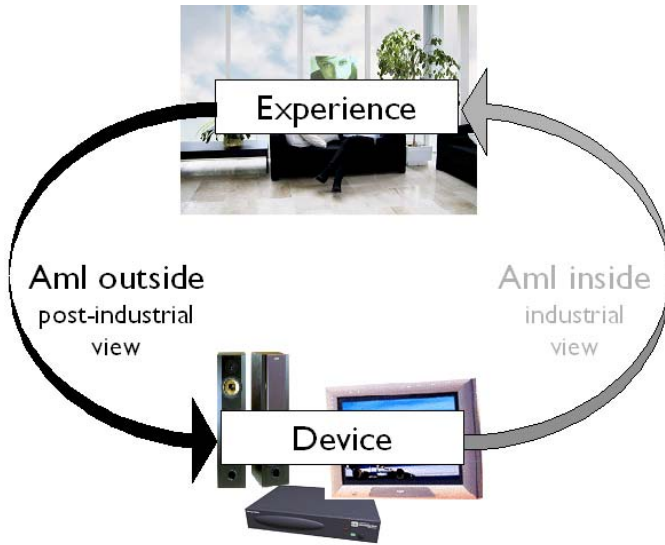
So the question becomes, how can we customize Ambient Intelligence environments on a mass basis to achieve economies of scale as well as scope?

To address this research question we iterate between a top-down and bottom-up approach. There are several reasons for this. Over the past years, many Ambient Intelligence prototypes and scenarios have been developed and written. Unfortunately, there are so many examples and there is so much variety in these examples that it is very hard to generalize from these examples. There is also no clear definition of what Ambient Intelligence is or how it relates to other forms of media. In the first part of this article we therefore examine important social, economical, and cultural trends. We then propose the notion of *ambient narratives* as an organizing concept. In the second part, we discuss the design and implementation of an ambient narrative system in detail and report on the results we obtained from experiments.

## Part I

### 2 Disembodied Intelligence

Figure 1 introduces two high-level strategies towards Ambient Intelligence. In the *Ambient Intelligence inside* approach the application intelligence is embedded in a tangible, physical product either in hardware or as embedded software on hardware. This strategy is a product-centric approach: The device generates the experience. The *Ambient Intelligence outside* approach starts from the desired experience: Devices are the means through which the desired experience is conveyed to the end-user. The application intelligence is disembodied from the tangible, physical product and offered as an information product or service to the end-user. In both cases the application intelligence can be personalized to the end-user. The differences between the two approaches become clear if we look at the underlying economics.



**Fig. 1.** Two different strategies

The product-centric Ambient Intelligence inside approach is the dominant, industrial strategy towards realizing the ambient intelligence vision. Market studies and surveys are conducted to define products that are targeted to specific customer segments. Products are mass-produced and delivered via large distribution and retail networks to the customer. The focus of companies that follow this Ambient Intelligence inside approach is on achieving product leadership in one or more product categories. With their industrial mindset, these companies are often forced to view Ambient Intelligence as a new and complex physical product or range of cooperating products that are mass produced and brought to the market by means of mass marketing and mass production techniques.

This mode of operation is increasingly under pressure. Over the past decades the needs and wants of individual people have become less and less homogeneous and stable, and, as a result of high-speed changes in society and technology, the pace of life also greatly increased; our relationships with people, places, things, organizations and ideas have shortened in duration in the past decades [39]. Power is also shifting from companies to individuals, today customers are no longer at the end of the value chain; they are at the beginning. In such a rapid-changing and demand-driven society, customers demand personalized products and services at any time so producers who can customize their products and services in real time will have a decisive advantage over producers who cannot deliver in real time [19]. Physical product manufacturers are constantly trying to reduce time-to-market to deal with this reality, but eventually only information products that have zero mass and do not occupy space can be custom produced and delivered in real time. With products and services becoming commoditized quickly, innovation becomes all the more important. But as the speed of innovation can be much higher for information goods than for physical products, product innovation is easier if the product or service is informational or as much informational as possible (as not all products can be turned into digital information goods).

In this context, the demand-driven, informational Ambient Intelligence outside strategy is a better fit than the supply-driven and product-centric Ambient Intelligence inside strategy. This leads us to a first requirement: *Ambient Intelligence is a personalized information good delivered over a set of networked devices which surround the customer(s). The intelligence that controls the appliances is separated from the ambience and customized on-demand by individual customers.*

### 3 Experience Co-creation

The rising affluence and growing transience of people, places, things and ideas that created the material basis for today's global, informational networked society led to another, even more profound change. This change is however more subtle and occurs deep inside people's minds but affects the economy at large: People's existential view towards life is shifting from purely external oriented to more and more internal oriented [33]. This difference is best explained by an example. If we are external oriented, we focus on the effect on the outside world. Driving a car to go from home to work is an example. If we act in an internal oriented way, the effect on ourselves, the resulting experience is important: We want a good driving sensation. This situation management with the goal to affect one's own inner life is a crucial change from the past in which most people were mostly concerned with survival or acquiring higher social (economical) status. Life may seem easy if everything is available in abundance, but the difficulty of biological survival is replaced with the question of "What do I want next?". To support people in this need companies are changing their focus from manufacturing products and delivering services to staging memorable experiences and transformations.

The actors in the experience economy, experience consumers and producers, each have their own rationality type, i.e. a collection of action strategies with constant, repeating goals. The consumer acts in a purely internal oriented way: Experience consumers actively work to regulate the constant stream of experiences they experience. The experience consumer rationality can be summarized by the following principles [33]: Correspondence, abstraction, accumulation, variation and autosuggestion. The correspondence principle states that people do not choose an experience at random, but that they choose experiences that they connect with in their own style. Abstraction means that experience consumers do not look for the optimization of a single experience offering but the optimization of a flow of experience offerings. Accumulation refers to the fact that people have the tendency to increase the tempo in which they consume experiences and start to collect, and pile, experiences. To still feel a nice stimulus after each accumulated experience, people look also for variation. This variation is often restricted to a common frame: e.g. people change bars but not do change the bar scene. Autosuggestion finally refers to the fact that people have a strong need to ask other people in the same environment what they felt of a particular experience being offered in the absence of any clear quality criteria to measure the value of an experience. In stark contrast to the internal oriented rationality type of the consumer is the external oriented rationality type of the experience producer. The Experience producer tries to match the action patterns of the experience consumers as best as possible with their strategies. They have developed a strong (traditional) outside oriented rationality type. Their rationality type can be characterized with the following action strategies: Schematization, profiling, variation and suggestion. Schematization refers to the fact that demand for experience can be categorized in aesthetic schemas (coherent style groups), which are common in large groups of the population and stay stable over the years. Segmentation of the market according to aesthetic schemas alone is not sufficient, companies must also profile themselves to make themselves known by means of elaborate brand management campaigns (profiling). Since experience consumers demand variety and expect the experience to have something new, something more stimulating, the experience producer must be able to customize the experience on a mass, low-cost basis (variation). Finally, providers of experiences must work hard to raise the suggestion that their product is new or authentic (suggestion).

To better understand what experiences are, we can look at the behavior of actors, consumers and producers in the experience economy and society at large or look at the formation of individual experiences. Eventually, the behavior of the actors in the experience economy at large must emerge from how individuals form experiences in their mind. We can view the individual as a free subject that interacts with its environment, or as an object that is exposed to environments that impose meaning. In reality both views are true at the same time as each individual is both subject and object at the same time. Experiences can therefore be said to form in a subject-determined, reflexive and involuntary way. The analysis of how experiences are formed can be approached from two different angles:

- The subject-oriented viewpoint puts the performance of the self in the foreground. Performance is the main object of study.
- The object-oriented viewpoint focuses on the production of signs and meanings that are encoded in for example a written or audio-visual language by an author. The main object of study is text (in the broad sense of the word as we will explain).

### 3.1 Subject-Oriented

The word performing is so commonly used in our language that we forget that performances are pervasive to every culture. Performances are not just seen on in the theater, music, dance and performance arts in general but also in our everyday lives: We perform the role of a father or son in our private lives but maybe also that of a doctor, judge or police agent for example in our professions. Performance studies is an emerging yet wide-ranging interdisciplinary field that takes performance as the organizing concept for the study of a wide range of human behavior. It embraces research in the social sciences from anthropology to theatrical studies. Because performances vary so widely from medium to medium and culture to culture, it is hard to pin down an exact definition for performance. Schechner defines performance as ‘ritualized behavior conditioned/permeated by play’ or ‘twice-behaved behavior’ [32]. When people are performing, they show behavior that is at least practiced once before in a similar manner. In traditional performance arts this behavior can be detected easily: Actors in a theater play, opera or movie rehearse their roles off-stage and repeat this behavior when they are on stage. But this twice-behaved behavior can also be seen in a priest conducting a wedding ceremony, a surgeon operating on a patient or a McDonalds service employee behind the counter. Even in our own homes, people show signs of this repeated behavior. This happens for example during everyday rituals, like brushing your teeth in front of a mirror in the morning, watching a soccer match with friends, or, coming home from work in the evening. Note that, here, the sending and receiving party in a performance may be the same. These kinds of everyday life performances were already mentioned by Goffman [14], who wrote in his book the ‘Presentation of Self in Everyday Life’ about how people follow culturally specified social scripts that interact with each other. These social scripts may differ from culture to culture and from epoch to epoch, but, according to Goffman, no culture exists without social scripts. People are performing all the time, most of the time without knowing. Each time we perform one of these social scripts, we form a new experience. These new experiences can also be shared with everybody else present, as in interactive theatre or live action role playing games. Here the audience actively participates in the performance or play, following certain predefined/rehearsed roles and rules that are combined to create a novel experience each time the play is performed. The experience is co-created by the actors and the audience. This immediacy of the action creates highly immersive and engaging experiences [12].

Viewing life as a social theatre is interesting for us for two reasons: First, if people behave according to social scripts, we may succeed in *codifying interactive*

*media applications to support people in carrying out these scripts.* Just as lighting and sound effects add to the overall drama of a theatre play, Ambient Intelligence may thus be applied to enhance the performance described by these social scripts. Second, positioning Ambient Intelligence in performance theory may open up a well-studied and familiar frame of reference for the design of Ambient Intelligence environments and the underlying technology, as we will discuss later.

Many social scripts are enacted at home. The home people live in can be seen as a stage on which perform everyday rituals, such as brushing your teeth, going to work or watching a soccer match. Figure 2 shows a cartoon projected on a Philips mirror TV (a two-way mirror with an LCD screen behind) that invites the small child standing in front of the mirror to brush his teeth for two minutes. The cartoon carries the child through this daily task. It would be too easy to say that we can create an optimal experience by making a task more effective or more entertaining. A better characterisation of an optimal experience is perhaps provided by psychologist Mihaly Csikszentmihalyi, who argues that happiness is not so much a result of finishing a task but more about being immersed and engaged in the process of performing the task. Only then do we get into a state of ‘flow’ and optimal experience [8]. In the mirror TV example, the cartoon shifts the attention of the child from achieving the end result to the process of getting there. The cartoon increases the flow of the activity by improving the level of engagement.



**Fig. 2.** Enhancing everyday life performances with Ambient Intelligence

The central role of performances is also reflected in recent business literature about services. Pine and Gillmore [28] talk about an experience economy in which work is theatre and every business a stage. Earlier research on service marketing by Fisk and Grove [13] discusses a theatre framework for service marketing, in which services are seen as performances, front-end service personnel as actors, the service setting as the stage on which the service is performed, products used in the services as props and the business process of the service as the script. Empirical evidence suggests that the ‘servicescape’, the environment of the service, plays an important role in how people perceive a service encounter. This suggests that Ambient Intelligence can also be applied in professional service encounters to enhance the atmospherics of a service encounter and thereby the perception or performance of a service in a positive way.

Consider for example a medical imaging room in a hospital. Many patients feel frightened by the bulky equipment in the examination room of a hospital.



By enhancing this environment with immersive media, e.g. by projecting video clips on the walls and ceiling of the examination room, patients may feel more at ease, as illustrated in Figure 3. Philips Medical Systems demonstrated this concept together with Philips Design in the ambient experience pavilion at the Annual Meeting of the Radiological Society of North America (RSNA) in 2003 (see Figure 3).



**Fig. 3.** Enhancing a medical examination room with Ambient Intelligence

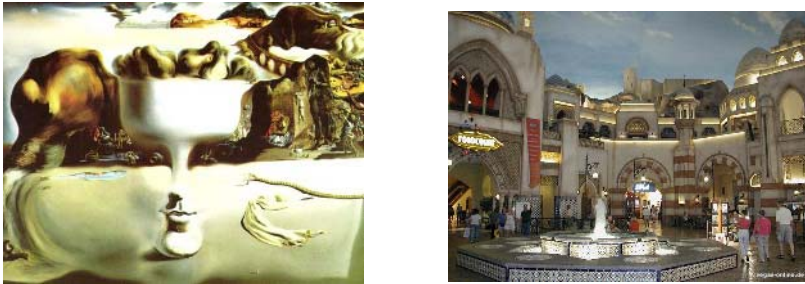
### 3.2 Object-Oriented

To understand how experiences are formed, we can also look at how meaning is imposed upon individuals. Literature studies can offer insight in this process [29]. Some schools argue that this meaning is entirely contained in the structure of the text (e.g. structuralism, Russian formalism), others such as Marxism, feminism, gay/lesbian studies argue that the social context is vital to understand this. Other theories of literature such as reader response theory take a much more subject-oriented view and consider the reader in the analysis of the text. Nobody reads the same text. Nobody interprets the meaning in the text in the same way, this depends on the experience of the individual reader.

In most works of literature and therefore literary analysis the text may be read differently by different readers, but the text to which readers are exposed is always the same. In interactive fiction, readers can actively change, affect and alter the plot of the narrative. Although works of interactive fiction and narrative are known from earlier times, interactive fiction became easier and popular with the advent of digital computers. In ‘Hamlet on the Holodeck’, Janet Murray [27] argues that the computer is not just a tool, but, a *medium for telling stories*. Since ‘Adventure’ the first text adventure, the computer adventure game genre has evolved and become much more graphical and integrated often in first person action games but the reader that can affect or otherwise alter the plot remains present. The non-linear nature of interactive narratives also attracted the attention of hypertext writers. Michael Joyces ‘Afternoon, a story’ (1987) and Stuart Moulthrop’s ‘Victory Garden’ (1991) are some successful examples of hypertext novels. In interactive narratives, the reader actively transforms in a performer but the performance is changed, affected and altered by the text.

## 4 Ambient Narratives

Interactive narrative and interactive drama cannot be understood from only the object-oriented or subject-oriented lens. The analysis depends on the choices of the subject but at the same time the choices of the subject are constrained, influenced and maybe even unconsciously or consciously directed by the text. This tension between reader interactivity and plot structure is present in all works of interactive narrative and drama. Too much interaction and the reader/performer feels lost. Too little interaction and a reader/performer feels his actions do not matter. Believability is another important factor [5] in interactive drama (as well as in human computer interaction design in general [21]). The characters in the story world may seem alien, but their actions should be believable at all times within the context of the story world in which they live. Murray [27] proposes three categories for the analysis of interactive story experiences: immersion, agency and transformation. Immersion is the feeling of being present in another place and part of the action, it relates strongly to believability. Agency is the feeling of empowerment that comes from being able to take actions that relate to the players intention. Transformation is the effect that happens if a players self is altered by the multitude of possible perspectives offered by an interactive story experience.



**Fig. 4.** Different forms of interactive narrative

Forms of interactive narrative are not just found in hypertext fiction or computer games, they can be found in many, perhaps unexpected, places. Painters and sculptors use perspective to tell different stories in one painting or sculpture. Depending on the way you look at the work of art, a different story is revealed to you. Figure 4 (left) shows the famous painting ‘Appartition Of A Face & Fruit Dish’ by Spanish surrealist painter and artist Salvador Dali. Depending on the way you look at this painting you can see a dog, a cup with hanging fruit, a face and a desert landscape with the sea in the background.

Interactive narrative is also always present in architecture. Depending on the way you walk through a building a different story is told to you. This becomes very clear in museums but also is present in other forms of architecture. Theme parks and themed environments such as the Desert Passage in Las Vegas (see

right-hand side of Figure 4) are explicitly designed to support interactive storytelling.

If we ‘add’ up all such single interactive narratives, we get a single master narrative that might be called an *ambient narrative*. An ambient narrative is superposition of the interactive narrative forms found in architecture, computer games and virtual reality, and works of art. An ambient narrative has the following characteristics:

- Interactive: Readers create their own unique, personal story and experience from a large set of possible choices defined by the author of the interactive narrative.
- Situated in mixed reality: Most forms of interactive narratives and drama are situated in the real world (architecture, improvisational theatre, visual art) or in the virtual world (e.g. virtual reality simulations and adventure games). Ambient narratives have both a real world and a virtual component. ‘Text’ and ‘reading’ should therefore be taken broadly: Reading becomes performing in such a mixed reality environment; we may skip a text page as we move from one room into the next. In contrast to augmented reality applications (see e.g. [4]), users do not necessarily need special equipment such as head-mounted displays or see-through glasses to experience these virtual objects. The virtual objects are presented through the individual devices that surround users. Each device generates only a part of the final experience.
- Non-fictional: Ambient narratives can be designed for pure entertainment or infotainment but will be mostly designed to support people in their everyday life activities and rituals (e.g. brushing teeth, cooking etc.).

How does this relate to Ambient Intelligence? By making choices in the ambient narrative we choose our situation (subject-determined), but this situation will affect, alter us as objects to the meaning imposed on us by the ambient narrative in an involuntary way (object-determined). For us as performing readers in the ambient narrative, Ambient Intelligence is a subset of the overall story and the experience formed in our mind. More precisely, Ambient Intelligence is that part of the co-created story that is conveyed through the user interface of the devices surrounding us.

The notion of ambient narratives brings us the organizing concept we are looking for because ambient narratives support the mass customization of ambient intelligent environments. This organizing concept is not built on quicksand. This definition of ambient narrative helps to understand the importance and pervasiveness of interactive texts and performances. We can start to view Ambient Intelligence as something more than a technological phenomenon; we can start to see it as a by-product of a literary form and apply literary theory and performance theory to deepen our understanding. The ambient narrative can be written completely in advance by an experience designer, the author of the ambient narrative like a computer adventure game or only partially, more like a live action role playing game or piece of improvisational theatre. In both cases, the ambient narrative allows readers to make their own choices that affect the

plot. In both cases Ambient Intelligence is a by-product of a co-creation process between reader and writer. The difference is that in the latter case, the reader may also alter the plot at a more fundamental level by adding his own material to the ambient narrative: The ambient narrative is not static, but dynamically evolving.

We can also derive a more concrete but weaker definition of ambient narratives that helps computer scientists, system architects and software engineers to model media-enhanced performances in the home and in commercial service encounters in a machine understandable way. This definition will be introduced in the second part and forms the basis for the design and implementation of the ambient narrative prototype and simulation tool that is under development. First, we give a scenario of a typically ambient intelligence scenario that will be used as an example throughout the second part.

## Part II

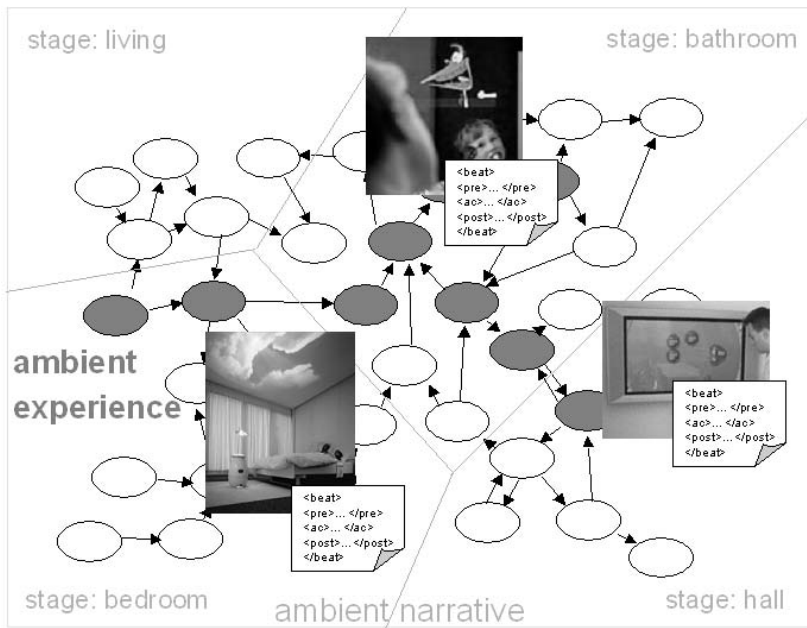
### 5 Example Scenario

Sarah, age 35, married to John arrives at home after a long working day. When she enters the hall, the lights automatically switch on and in the mirror TV in the hall she sees her voice mails and text messages. She quickly browses through them and then leaves to the kitchen to prepare some dinner. When she enters the kitchen, the kitchen portal is shown on the screen in the kitchen. She decides to start an interactive, electronic cooking book application on this screen. After a short while her attention falls on a simple but very nice looking Italian pasta dish which she can prepare with the ingredients present in the house. As soon as she has selected the pasta dish, a voice controlled application guides her through all the necessary steps to prepare the food. Eventually, when the food is ready to be served, she selects a video chat application that opens up a communications channel in each room where a person is. John, who just entered the hall, sees Sarah and responds that he will be right there. Sarah takes the pasta dish to the dining room and John sets the lights, wall-size displays and ambient music to an Italian restaurant.

### 6 Ubiquitous Physical Hypermedia

An ambient narrative can be represented by a hypertext network with nodes and links between nodes. This is illustrated in Figure 5. Each node is a description of an interactive, distributed media application in a ubiquitous, physical hypermedia language that corresponds with a particular everyday performance, for example brushing teeth or a medical examination in a hospital performed by a doctor. Links between nodes model allowed transitions from one media-enhanced performance to another. Transitions should be taken very broadly in physical hypertext; walking from one room to another, changing orientation in

a room, touching a lamp can all be seen as events that activate a link. By making choices in the hypertext network, the reader creates a personal story – a personal ambient intelligence experience. Ambient narratives enable mass customized ambient intelligence; the main difference with mass customization in the traditional sense is that the mass customization process in ambient narratives is continuous as opposed to the assembly of personal computers or design of furniture. When readers start to add and remove their own plot material to the ambient narrative, the mass customization process turns into a co-creation process where readers and writers create the desired experience together.



**Fig. 5.** An ambient narrative as a hypertext network

From a writer's perspective, the ambient narrative describes all possible media-enhanced performances and their interrelationships. In a pure mass customization approach, the complete ambient narrative is written in advance, but still enables consumers to create their own personal story, their own Ambient Intelligence, from existing plot material. In our example, this means that ambient narrative authors should understand the performances, rituals that take place in Sarah's home. Although this seems a daunting task, there are a number of factors that simplify this task: First, many performances are framed by location. In our example, we know beforehand that Sarah is not entering the kitchen to take a shower. Furthermore, even if people may now have more choice than ever before, mass media and popular culture tend to reinforce cultural identity and

form new social groups. Authors of ambient narratives could take this into account and design a variety of ambient narratives, each suited for a different life style. Finally, like in movies and drama, people who interact with an ambient narrative may be willing to accept errors in the performance as long as they do not disrupt the overall experience too much.

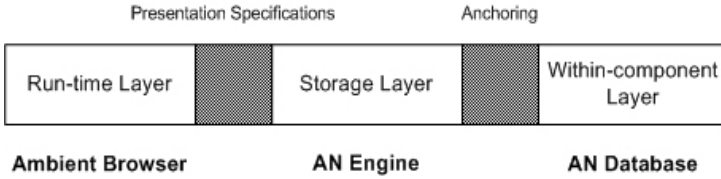
In a co-creation strategy, the ambient narrative plot material is not fully defined in advance but (implicitly) created by the readers of the ambient narrative. This allows end-users to program their own ubiquitous computing environment in exactly the way they see fit. Unless explicitly mentioned, we will look at ambient narratives from the writer's point of view, as our focus is on the production of Ambient Intelligence.

## 7 Related Work

Ambient narratives can be related to interactive storytelling, hypertext, and ubiquitous computing.

Many interactive storytelling systems designed so far in the interactive fiction domain are based on Artificial Intelligence techniques that control either the plot generation or real-time character behavior, see e.g. [5, 24, 23]. Because AI planning techniques deal with action selection, these techniques lend themselves also well to sequence plot elements into a compelling dramatic story. One of the main issues in interactive drama systems is the balance between player interactivity and plot structure: Too much structure imposed, and users may get the impression that their actions do not matter, because the important choices have been made anyway. Too much focus on interactivity on the other hand causes the impression of feeling 'lost'. Believability is another important factor [5] in interactive drama (as well as in human computer interaction design in general [21]). The characters in the story world may seem alien, but their actions should be believable at all times within the context of the story world in which they live. Murray [27] proposes three categories for the analysis of interactive story experiences: immersion, agency and transformation. Immersion is the feeling of being present in another place and part of the action, it relates strongly to believability. Agency is the feeling of empowerment that comes from being able to take actions that relate to the player's intention. Transformation is the effect that happens if a player's self is altered by the multitude of possible perspectives offered by an interactive story experience.

The non-linear nature of interactive narratives also attracted the attention of both hypertext writers and researchers. Michael Joyce's 'Afternoon, a story' (1987) and Stuart Moulthrop's 'Victory Garden' (1991) are some successful examples of hypertext novels. During the nineties, hypermedia narratives were also demonstrated, HyperSpeech [3] and HyperCafe [31] to name just a few. Recent publications have expanded models and techniques into mixed reality environments [15, 34, 30]. With respect to interactive storytelling, Weal et al. [40] describe an approach that records the activities of children in an outdoor environment and represent these recorded activities using adaptive hypermedia



**Fig. 6.** Mapping the high-level system architecture onto the Dexter reference model

techniques to children later so they can later relive the experience. Romero and Correia [30] discuss a hypermedia model for mixed reality and describe a mixed reality game that takes place in a gallery environment.

Pervasive or ubiquitous computing plays an important role in ambient narratives. To create the feeling of ambient intelligence, users need to be surrounded by devices that can work in concert on behalf of the user. The term ubiquitous computing was coined by Mark Weiser [41] in the early 90s and many work has been done since. It is outside our scope to discuss this work in detail, see [2] for an overview of applications and research topics such as wireless networking, context-awareness, power management etc.

## 8 Dexter-Based Hypertext Model

The ambient narrative approach is modular, because the narrative itself constitutes the modular parts of all the possible Ambient Intelligence fragments. The user interacts with the narrative engine by the *ambient browser*, that collects user input and context information from multiple devices. The *ambient narrative engine* determines the next episode in the ‘story’ told, given the user input and context and the current position in the ambient narrative plot. It returns a description of the next episode that is rendered by the ambient browser. Expressed in business terminology we follow the paradigm of mass-customisation [9], where the ambient browser and ambient narrative engine assemble a customized product, Ambient Intelligence, and deliver it to the customer, the reader/performer.

The remainder of this section describes how the high-level ambient narrative system architecture above can be implemented in terms of (extensions of) the Dexter hypertext reference model[16]. Figure 6 shows the mapping of our high-level ambient narrative system architecture on the Dexter model. The following subsections explain how the run-time layer is implemented by the ambient browser, the storage layer by the ambient narrative engine and the within-component layer by the plot material of an ambient narrative. In the following sections, we will discuss the current system architecture and the results of experiments conducted with an ambient narrative simulation tool we have written.

### 8.1 Run-Time Layer: Ambient Browser

Although hypermedia can be added to the Dexter model as a data type in its storage layer, this approach cannot adequately support the complex temporal rela-

tionships among data items, high-level presentation attributes and link context that are important in supporting hypermedia. The Amsterdam Hypermedia Model (AHM) [17] extends the Dexter model by adding these hypermedia requirements. The AHM model inspired the definition of the Synchronized Multimedia Integration Language (SMIL) [37]. AHM and the SMIL language are designed for hypermedia presentations running on a single device and therefore do not mention the issue of timing and synchronization of media elements across multiple devices, characteristic for mixed reality or ubiquitous hypermedia. To support timing and synchronization of media objects within and across devices, we use an in-house developed SMIL interpreter. This SMIL interpreter has the role of a networked service to which the other devices register themselves. The underlying reason for this choice is that we expect ‘media’ to also include lamps, fans and other output devices in the future. To use this functionality in SMIL, we have extended the toplevel element in the SMIL language with a proprietary ‘target’ attribute that specifies the rendering (or input) device. The author of the SMIL document can set the target attribute of the toplevel element in the SMIL head part to point to a specific rendering or input device. In the SMIL body part, the author can use (as he would normally do) the id of the toplevel element or the id of one of its region element children in the region attribute of a media element (e.g., an image, or a video fragment), to indicate the device/region on which the media element should be rendered. The advantage of this approach is that we can remain close to the AHM model and do not have to introduce spatial mapping functions outside the SMIL engine as described in [20] for example.

Since every performance can be augmented with a hypermedia document and multiple performances can be going on at the same moment in an ambient intelligent environment, we have extended the run-time layer further to deal with document sets. Users can add, remove or replace documents from the set of documents in the ambient browser and in doing so change parts of the surrounding ambience.

From a high-level point of view, the presentation and interaction with SMIL belongs to the run-time layer in the Dexter model. This view differs to some extent from the role of media objects in the AHM, which addresses mostly the storage layer. We consider SMIL documents as basic building blocks, i.e., the components in the storage layer. The AHM model however views individual media objects as the principal components. The advantage of our approach is that we can abstract from the low-level positioning of media objects in space and time and focus on describing how such hypermedia presentations are selected based on context situations. This does not mean that the AHM model is not present; it is just hidden. Inside the SMIL components, the AHM model is revealed. Since this extended SMIL document is viewed as a component, we could replace it by a different hypermedia markup language, or use dedicated applications with new forms of multimodal interaction (e.g., those not supported by the SMIL implementation).

## 8.2 Storage Layer: Ambient Narrative Navigation

The storage layer in the Dexter model describes the mechanisms by which nodes and links are connected to form networks. The storage layer itself is not



specific about the internal structure of the components; components are treated as generic data containers. The components of the ambient narrative are called *beats*. The term beat originates from theater play and film script-writing and is defined [25] as a change in the behavior of a character. The beat is the smallest unit within a scene, that in turn represents a greater change in behavior. Sequences of scenes form acts which culminate into even greater change. This beat (sequencing) approach is also taken in interactive drama, e.g. by Mateas [24], Magerko [23] and Brooks [6]. Since ambient narratives are also enacted, we choose the same terminology. So, an ambient narrative can be viewed equivalent to a hypertext network consisting of beats (nodes) and links between these beats.

An individual Ambient Intelligence experience is represented by a sequence of beats and any specific parameters that may have been used in the instantiation of these beats. Like in Schank's script theory, specific memories (in our case Ambient Intelligence experiences) are represented as pointers to generalized episodes (beats) plus any unique events for a particular episode (represented by story values, as explained shortly).

**Beat Language.** Now, we discuss first a language for representing beats, followed by an explanation of how beats are sequenced by a beat sequencing engine or ambient narrative engine. We analyzed dozens of Ambient Intelligence scenarios from literature and past projects in our groups and derived the beat markup language and beat sequencing algorithms from this analysis. Here we will explain the design choices made, illustrated with examples taken from the scenario of section 5.

In mixed reality environments, context and text belong together. Presentation descriptions should only be selected if both the text (position in the narrative) and the context match the current situation (i.e., position in the real-world). Therefore, each beat description consists of two parts: a *preconditions* part to represent the context, and an *action* part to represent the 'text'.

The beat sequencing engine (or ambient narrative engine) that controls the navigation through the ambient narrative checks the *preconditions* part during the selection or search for new beat components. It specifies the conditions that must hold before the beat can be selected; restrictions can be set on the performance (activity), actors (users) that need to be present, stage (location), props (tangible objects and devices present) and/or script (session variables).

The *action* part is executed after the beat has been selected by the engine for sequencing. It is subdivided in an initialization part, a main part and a finalization part. The initialization part is executed when the beat is activated and before the presentation description in the main part is executed. The main part contains the actual presentation description, encoded by a ubiquitous hypermedia document. The finalization part can contain instructions that specify what should happen if the beat is removed.

Because one beat alone will not be perceived as ambient intelligence by readers/performers of the ambient narrative, a requirement for the beat markup language is that it supports links between beats or queries for beats that can

be activated by readers. For example, when Sarah decides to go for the Italian dish, she explicitly starts the prepare food application that helps her through the cooking performance.

Clearly, it is too restrictive to support only one activity at a time: At any time there are typically  $N$  actors involved in  $M$  performances in such an environment. To model multiple activities in the model, it is necessary to support concurrent browsing paths in the model. We looked at a number of approaches discussed in literature. The Trellis system explained in [36] uses a formal Petri net based model to model synchronization of simultaneous traversals of separate paths through a hypertext. The Petri net based model probably works best if the hypertext network is not open-ended, but we would like to support that both authors and end-users can add/remove beats from the narrative as well as allow conditional links (queries) between beats to make room for more variety and surprise during navigation. The spreading activation network introduced in [22] could also be a candidate to model concurrent browsing if we treat a beat as a competence module of an autonomous agent (the ambient narrative engine) and model the action selection (beat selection) as an emergent property of an activation/inhibition dynamics among these modules. We decided to limit ourselves to a more straightforward solution however, and simply add a ‘behavior attribute’ to the link that indicates if the target of the link (query on the beat database) should be added to the existing set and/or if the source of the link (query on the active beat set) should be removed from the existing set.

Beat links and queries are traversed immediately as a response to a specific user event (e.g. press of a mouse button, touching a part of the screen, stepping on a floor sensor). In many Ambient Intelligence scenarios, applications are not triggered immediately but as soon as certain context conditions have been fulfilled. We define a context trigger in the beat language as an element that has a preconditions part and a link. When the trigger is set, the system waits for the preconditions specified in the trigger(s) to become true. If all preconditions are valid, the system traverses the link. The situation where Sarah enters the hall and the voice and text messages are presented on the screen in the hall is modelled as a context trigger.

For many Ambient Intelligence applications in the scenarios we investigated it proved too inflexible to allow for only implicit context changes and explicit user feedback. The choice of the next beat can also be dependent on information from the past (e.g. user preferences), present (current narrative state) or future (user goals set). We want to prevent that Sarah can select the food preparing application if she has not selected a recipe first in the electronic recipe book for example. A further requirement for the beat language is therefore a mechanism to test and change variables. These variables that change during the session are called *story values* and are stored in a global *story memory* (see below). The author of an ambient narrative can specify in the precondition part of a beat or trigger which story values must hold. In the action part of a beat the author can define story value changes. These story value changes will be written in the story memory and influence the selection of new beats. At the moment, each

story value is a simple key value pair and the story memory is simply the list of all story values that have been set.

Links, triggers and story-value changes elements can be placed in either the initialization, main and/or finalization sections in a beat to describe possible transitions between beats. These elements are called from the run-time layer.

Table 1 summarizes the main beat language features.

**Table 1.** Main beat language requirements

| <i>Requirement</i>                  | <i>Language Construct</i>             |
|-------------------------------------|---------------------------------------|
| context and text description joined | beat                                  |
| user-based navigation               | link                                  |
| simultaneous performances           | beat sets                             |
| context-based navigation            | trigger                               |
| session variables                   | story values stored in a story memory |

The following XML fragment is a typical beat description. It contains a `pre` element that describes the minimal required context conditions and an action part that describes what happens when the beat has been scheduled for execution. For space reasons the presentation markup has been omitted.

```
<!-- OnBrowsingRecipes:

Starts an application that shows electronic recipes the user can choose and customize.
The OnBrowsingRecipes, OnPreparingFood and OnCommunicating beats model a workflow.

-->
<!DOCTYPE beat SYSTEM "http://localhost:8080/beatdb/HomeLab/beat.dtd">
<beat id="Kitchen_OnBrowsingRecipes1">
  <!-- preconditions that must hold for this beat to become selected -->
  <pre>
    <stage id="Kitchen" time="day" location="HomeLab">
      <performance id="BrowsingRecipes">
        <actor id="housemate"/>
        <prop id="screen2" capability="SVGA touchScreen"/>
      </performance>
    </stage>
  </pre>
  <!-- what happens when the beat is selected -->
  <action>
    <init/>
    <main preview="
      http://localhost:8080/beatdb/HomeLab/previews/Kitchen_OnBrowsingRecipes1.jpg">
      <!-- presentation markup interleaved with beat instructions -->

      <!-- set a goal story value and query -->
      <story-value action="add" id="setPrepareFood" name="prepareFood" value="true"/>
      <link id="prepareFood" behavior="add" to=
        "http://localhost:8080/beatdb/HomeLab/queries/Kitchen_OnPreparingFood.sql"/>

      <!-- kill beat -->
      <link id="kill" behavior="delete" from=
        "http://localhost:8080/beatdb/HomeLab/queries/Kitchen_OnBrowsingRecipes1.sql"/>
    </main>
    <final/>
  </action>
</beat>
```

The design of this beat language has similarities with the interactive drama approaches taken by Mateas and Magerko. Our preconditions part is similar to the selection knowledge stage in Mateas' Facade interactive drama architecture [24] that also uses tests on story memory values (and prior probability on beats). The action stage in Facade roughly corresponds to our action and post-conditions parts in the beat document. Magerko's IDA architecture [23] represents plots at the scene level and consists of five stages: initial state, required events, background knowledge, content constraints and temporal constraints. The initial state sets up the scene and may be similar to the initialization part in our action stage. The required events and background knowledge are comparable with our preconditions stage, while the content constraints that limit the binding of the variables used in the required events are similar to the embedded queries in the main action description part of our ambient narrative beat documents.

**Beat Sequencing.** Beats are sequenced together to create a personalised story in mixed reality, the result of which is what we have called Ambient Intelligence. The beat descriptions and beat sequencing engine or ambient narrative engine can be seen as an adaptive hypertext system. Brusilovsky [7] describes adaptive hypermedia as follows: *“By adaptive hypermedia systems we mean all hypertext and hypermedia systems which reflect some features of the user in the user model and apply this model to adapt various visible aspects of the system to the user. In other words the system should satisfy three criteria: it should be a hypertext or hypermedia system, it should have a user model and it should be able to adapt the hypermedia model using this model.”* The Adaptive Hypermedia Application Model (AHAM) [11] builds further upon this definition and tries to fit adaptive hypertext and hypermedia in the Dexter model. It defines a hypermedia application as consisting of a domain model, user model, teaching model and adaptive engine. The domain model describes how the information is structured in nodes and links. The user model describes the knowledge level of the user and also keeps a record of the nodes visited by the user in the past. The teaching model consists of learning rules (pedagogical rules) which define how the domain model interacts with the user model. The adaptive engine performs the actual adaptation.

The beat descriptions and beat sequencing engine can be quite easily described in terminology of the AHAM model. Beats and their interrelationships form the domain model. The user model is implicit in the story memory of the ambient narrative engine: The story memory contains session knowledge which can contain user preferences as we discussed before. The story memory dynamically evolves out of the continuous interaction between users and the ambient narrative. The teaching model is encoded in the beat descriptions. The action part allows the author to alter story values that can affect the selection of new beats and content items. The adaptive engine is the ambient narrative engine that sequences beats. The ambient narrative engine must implement an action selection mechanism as its main task is to find the next best beat. We implemented the beat sequencing planner using a structured information retrieval

approach. Queries for new beats are encoded in the action part of the beat description and may contain both fixed parameters and story values. Beat queries are never explicitly entered by the user, they are selected and filled in by the beat sequencing engine based on user input and information present in the story memory.

The advantage of this approach is that it allows us to introduce adaptation at different levels of the narrative like the Facade [24] and Hyperdoc [26] systems. If we use presentation templates instead of finalized presentation descriptions we can allow for adaptation at the subnode level, i.e., change the media objects in a SMIL fragment in the action part of the beat. Beat queries enable us to describe adaptation at the link level: The choice of a beat can be made context-dependent by using story values or context information in the beat queries. The use of beat preconditions and beat queries also allow us to easily add new beats and content without taking the system down. This way we can defer editing decisions by the narrative engine on both the node and subnode level until the moment they are played out. The same technique is used by the Automatist Storyteller system [10] for content items only. As a result authoring effort is lowered because the author does not have to explicitly sequence story elements into a finished story or rewrite the existing narrative when new material is added. This also provides the writer with the flexibility to add beats for specific devices that can be chosen if these devices are owned by the reader. Furthermore, authoring tools could assist readers in creating their own beats and inserting them in their beat collection.

### 8.3 Within-Component Layer: Plot Material

Depending on whether you look at the narrative level or at the hypermedia level, the presentation document is either an atomic component or a composite component. In the first case the component content in the within-component layer are for example SMIL documents or even dedicated applications, in the second case the component content are media objects, text, images, audio, video and other modalities such as lightscripts. All the component information and content is indexed and stored in an ambient narrative database for fast retrieval by the ambient narrative engine.

## 9 System Architecture and Algorithms

Figure 7 shows the architecture of the ambient narrative engine. User feedback, contextual data implicitly derived from sensors and the contents of the story memory together, determine the direction through the ambient narrative and how the selected beats are customized and sequenced.

First we discuss each component in Figure 7, then we illustrate how the engine works with an example from our scenario.

The *beat database* contains the beat documents that can be chosen. The *content database* stores all media elements and third-party applications that are needed for the beat documents in the beat database. The *context database* maintains a model of the ambient narrative physical context. This information is

needed to verify the actor, prop and stage preconditions for beats. The *story memory* is a list of story values that can be set by beats. All the beats that are currently active are placed in the *active beats* set. Context triggers set by beats are stored in the *active triggers* list.

The *internal event server* receives events from the ambient browser (incoming arrow). As can be seen in example beat, the action part contains language elements (link, story-value) that are not meant for presentation by the ambient browser but for navigation through the ambient narrative. When the ambient browser encounters such an element, it forwards the element id and corresponding document id to the internal event server of the ambient narrative engine. This component determines if the element id is a link, trigger or story-value change element. In case of a link type, the element is forwarded to the *beat scheduler*. In case of a trigger or story value change, either the active trigger set or story memory is updated.

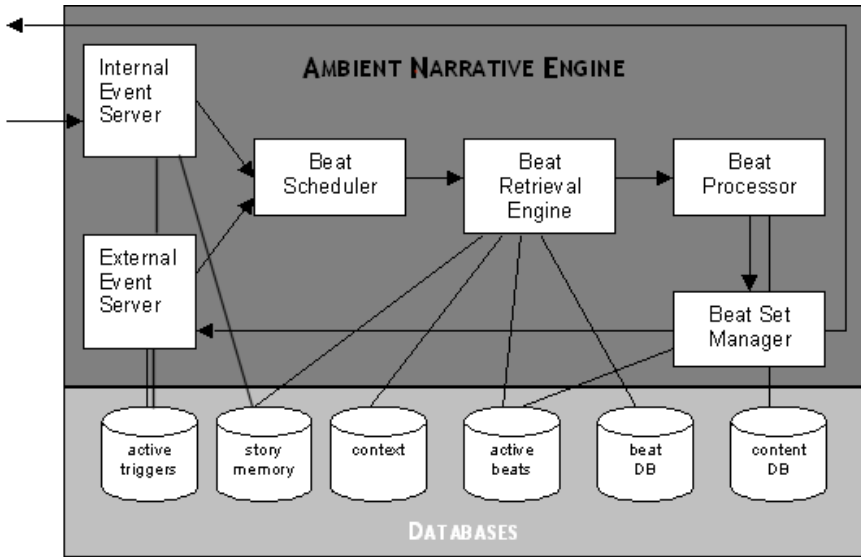
The beat scheduler may implement a scheduling and a filtering algorithm to change the order in which queries for beats arrive. In the current implementation the beat scheduler just passes the link information on to the *beat retrieval engine*.

The beat retrieval engine processes and executes the queries contained in the link elements: Each link consists of either a ‘to’ or ‘from’ attribute or both. The ‘to’ attribute contains a query for one or more new beats on the beat database, the ‘from’ attribute can be used to specify a query for removing beats from the active beat set. When a link element arrives, the beat retrieval engine executes the queries in the ‘to’ and/or ‘from’ fields and checks if the retrieved beat preconditions are fulfilled. Beats that are scheduled for removal are forwarded to the *beat set manager*. Beats that need to be added to the *beat set manager* are first processed by the *beat processor*. In our scenario, the example beat is added by the beat set manager as a result of a link that was triggered when Sarah entered the kitchen.

The beat processor parses the beat description and checks if the beat contains open parameters (adaptation at the sub-node level) for example embedded content queries that need to be filled in first before the beat can be send to the ambient browser. At this moment we do not support beat templates and assume beats are finalized at the moment of writing.

The beat set manager updates the active beat set and synchronizes this with the ambient browser to make sure the correct applications are running (those specified in the main part of the beats in the active beat set). The beat set manager also checks for any story value and/or trigger instructions in the initialization and finalization parts.

The *external event server* finally tests the preconditions of the beats in the active set and context trigger list. If a beat in the active beat set is no longer valid because one or more preconditions no longer hold, a number of different strategies are possible: remove the beat, look for a more specific or generic beat or do nothing. If all preconditions hold of one or more of the triggers in the context trigger list, the external event server will send the link that belong to the trigger to the beat scheduler for further processing.



**Fig. 7.** Ambient narrative engine architecture

When Sarah selects the video chat application to notify other people that the pasta dish is ready, the ambient narrative engine receives the element and document identifiers for two triggers (one for the hall and one for the living) from the ambient browser. The internal event server adds the triggers to the context trigger list and notifies the external event server to recompute the active beat set. The external event server checks the preconditions of the triggers by consulting the context database and story memory and finds that one of the triggers needs to be activated. The external event server then forwards the link to the beat scheduler for further processing. The beat scheduler decides to do nothing and sends the link to the beat retrieval engine which retrieves the beat that describes the video chat application in the hall. The beat retrieval engine forwards the resulting beat to the beat processor which checks if the beat description needs to be processed further, i.e. has any open parameters. The beat set manager adds the beat to the active beat set and notifies the ambient browser to update the running presentation. The result is that John sees a video chat application started in the hall with Sarah asking him to join dinner.

The design of the ambient narrative engine is inspired by the drama manager in the Facade system. This drama manager also uses user input and a story memory to sequence beats into an immersive interactive story. The difference with our approach is that we can not only adapt the link structure based on user interaction but also, to some extent, the nodes; for, we allow embedded queries in the beat description language. Similar concepts are also used by Magerko [23] and Szilas [38]: Magerko uses the word director while Szilas applies the term narrator to refer to a drama manager.

## 10 Experiments and Results

To verify the ambient narrative system, we developed an ambient narrative simulation tool. The reasons for building a simulation tool instead of a real prototype were two-fold: First, we wanted to reduce risks and development costs. Secondly, experience with building Ambient Intelligence type of applications and environment has taught us that it is hard to quickly prototype and debug interactive media environments. Therefore, we decided to work on a simulation tool first. Similar approaches towards rapid prototyping of context aware applications can be found in i.e. [35, 18].

The simulation tool fully implements the ambient narrative engine (storage layer), but emulates the behavior of the ambient browser (run-time layer) by means of a web interface through which users can manually enter events (user feedback and context changes) for the narrative engine and see the result of their action. The risk of this approach is of course that the simulation tool does not give a realistic experience, but since the ambient browser and narrative engine are relatively independent, we believe the risk is tolerable.

To test the concept, we picked fifteen Ambient Intelligence scenarios situated in the home (hall, kitchen and living) from the set of scenarios (130) we analyzed and wrote an ambient narrative that combined these scenarios. We evaluated the scenarios based on their complexity and difference to other scenarios in the set. This resulted in a network of 33 beats written in our beat markup language. We experimented with the ambient narrative by entering events directly in the simulation tool and viewing the effects on the active beat set. To make the simulation visually more appealing, we added a preview attribute to each beat. This preview attribute links to a picture or movie that illustrates how the beat would have looked like in a real ambient narrative system. During the simulation the ambient browser simulation shows the pictures that belong to the beats in the active beat set (see Figure 8).

Even with this relatively small ambient narrative, we can see interesting results and issues for further study.

Triggers, links and story-values create a highly dynamic system: For example, when a beat that has been launched by a context trigger is removed, it will be immediately reactivated if the context trigger is still set. Although this behavior may seem flawed at first instance, there are scenarios where this behavior is actually desired. For example, when a user logs off a public kiosk in a museum, the application should be restarted. In the experiment we found that we could realize the desired behavior simply by setting a story value that is tested in the precondition of the trigger or by removing the trigger. This same mechanism can be used for links: only traverse the link from the electronic recipe book to the preparing food application if story value ‘prepareFood’ has been set to true. Although it is possible to write the beats manually without any help of an authoring tool, authoring support is desirable for non-trivial ambient narratives because it can be quite difficult to figure out why a particular beat has been (re)activated.

In designing the ambient narrative documents, we also noticed that very dissimilar application scenarios share the same structure. For example the voice and



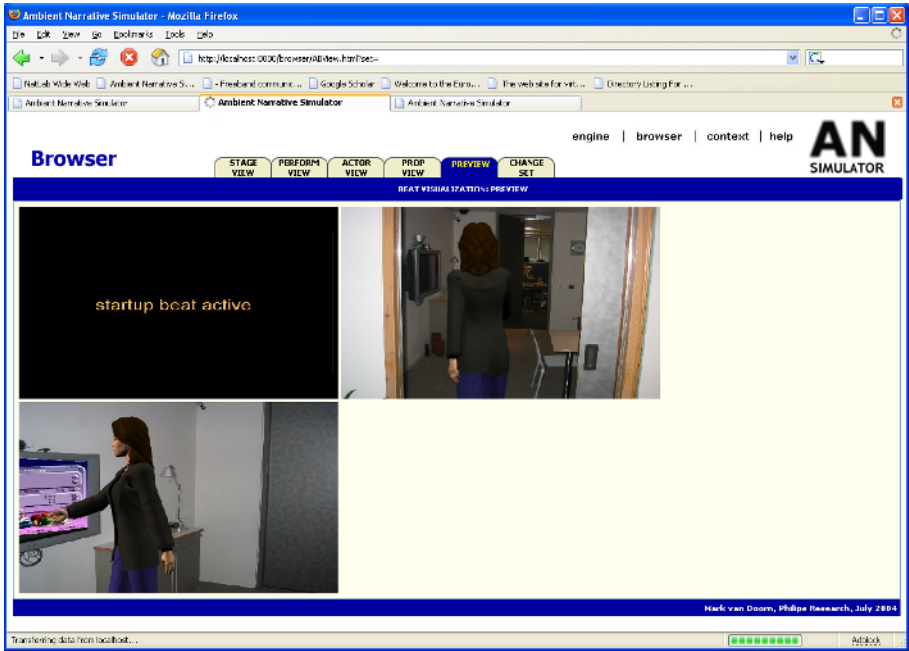


Fig. 8. Screenshot of Ambient Narrative Simulator

text messaging application in the hall and electronic recipe book application in the kitchen are structurally similar: On entering the room we set a context triggers that launches a portal type of application in the room from which readers can browse further. Workflow patterns like the selecting and preparation of food are modelled by setting and testing story values. Communication patterns are modelled using triggers: When Sarah selects the video chat application in the kitchen to notify Sarah the food is ready, triggers are set in the other rooms that test for the presence of a housemate. This pattern can be used for any modality and any number of presentations. We found that the function (beat relations) and context of beats is at least as important as the contents (interactive media presentation) of the beat contents themselves to understand the ambient narrative. This structuralist approach to understanding ambient narratives is just one angle of looking at the ambient narrative text. The interaction of users with the ambient narrative also deserves attention because ambient narratives are not really read but performed. The possibility of reusable plot structures for very different scenarios and domains may also simplify the development of an authoring tool, but this requires further research.

Tests with this example ambient narrative also revealed the need for some extensions on the model. At the moment we can model links between beats as a link or a database query. As a result there is no element of surprise or variation which leads to predictable behavior. While this may be sufficient (maybe even required) for functional, rational tasks, it is too restrictive for more emotional,

entertaining performances. Another issue we found while experimenting with the simulation tool is that users may want to save (part of) the current state of the ambient narrative session in their user profile. To support this, the model must support both persistent and non-persistent story values and the ability to update story values in the user profiles with the story memory and vice-versa. Furthermore, the localization model allow us to describe which props must be present on a stage but cannot express how props relate to each other: For example, it is possible to describe in our current beat markup language that a room should have a screen, surround loudspeakers, but it not possible to specify that the screen should be in front of the user and the loudspeaker around him at a particular maximum or minimum distance. The last issue we mention here is multiple user support. At the moment we can only describe (as part of the preconditions part of a beat or trigger) which actors must be on a stage and thereby who has access to beats. Relationships between actors and other attributes such as the minimum and maximum number of actors allowed in a performance are also required to model more complicated scenarios.

Next to design issues there are some implementation issues worth mentioning. The simulation tool user interface is adequate for testing all functionality, but requires manual input for all events and context changes which is inconvenient and can easily lead to typing errors. Early on in the implementation we decided to store all context and beat data in relational databases for algorithmic optimization and scalability reasons and develop the ambient narrative engine as a web application deployed on a J2EE-based web application platform.

## 11 Conclusions

Ambient Intelligence is a vision on the future of consumer electronics that refers to the presence of a digital environment that is sensitive, adaptive and responsive to the presence of people. Since it is technologically not possible to mass produce Ambient Intelligence with the current state of the art in artificial intelligence and economically not feasible to manually develop tailor-made Ambient Intelligence products or services for each customer individually, we believe a different approach is needed to move such type of ubiquitous computing environments out of research laboratories into the real world.

We described a mass customization strategy for Ambient Intelligence services offered over a collection of networked devices to customize Ambient Intelligence on a mass basis. In this approach, modular parts of Ambient Intelligence descriptions are assembled based on user feedback and interaction history into a personalized flow of personalized, interactive media presentations in which multiple devices may participate simultaneously. The Ambient Intelligence service thus allows people to create their own Ambient Intelligence within the scope of possibilities set down by the designer or writer of an ambient narrative, an interactive narrative in mixed reality that is designed to support people at home or in a particular service encounter in performing their everyday activities. We explained how an *ambient narrative system* can be implemented in the existing

Amsterdam Hypermedia Model (AHM) and Adaptive Hypermedia Application Model (AHAM). We implemented a prototype that simulates the concepts introduced in the setting of the home, and discussed the lessons we learned. Currently, we are working to connect the ambient narrative engine to a real device rendering platform to create a fully working prototype and test the combined system in HomeLab, the usability and feasibility lab at Philips Research.

Still, much work needs to be done and many improvements are possible on both the model and the algorithms used. This is why we decided to write this article in the first place. We hope that by introducing the ambient narrative concept and the hypertext model we have given you something to think about and perhaps even new ideas for AI algorithms and methods or new ways of applying existing AI algorithms and techniques in such ubiquitous computing environments.

## References

- [1] AARTS, E., AND MARZANO, S., Eds. *The New Everyday: Views on Ambient Intelligence*. 010 Publishers, 2003.
- [2] ABOWD, G., AND MYNATT, E. Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 1 (2000), 29–58.
- [3] ARONS, B. Hyperspeech: Navigating in Speech-Only Hypermedia. In *UK Conference on Hypertext* (1991), pp. 133–146.
- [4] AZUMA, R. A Survey of Augmented Reality. In *SIGGRAPH'95* (1995), pp. 1–38.
- [5] BATES, J., LOYALL, A., AND REILLY, W. Broad Agents. *Sigart Bulletin* 2, 4 (1991), 38–40.
- [6] BROOKS, K. *Metalinear Cinematic Narrative: Theory, Process and Tool*. PhD thesis, MIT Media Lab, 1999.
- [7] BRUSILOVSKY, P. Methods and Techniques of Adaptive Hypermedia. *User Modeling and User-Adapted Interaction* 6, 2-3 (1996), 87–129.
- [8] CSIKSZENTMIHALYI, M. *Flow: The Psychology of Optimal Experience*. Perennial, 1991.
- [9] DASILVEIRA, G., BORENSTEIN, D., AND FOGLIATTO, F. Mass Customization, Literature Review and Research Directions. *International Journal of Production Economics* 72, 1 (2001), 1–13.
- [10] DAVENPORT, G., AND MURTAUGH, M. Automatist Storyteller Systems and the Shifting Sands of Story. *IBM Systems Journal* 36, 3 (1997), 446–456.
- [11] DE BRA, P., HOUBEN, G.-J., AND WU, H. AHAM: A Dexter-Based Reference Model for Adaptive Hypermedia. In *UK Conference on Hypertext* (1999), pp. 147–156.
- [12] FALK, J. *Funology: From Usability to Enjoyment*. Kluwer Academic Publishers, 2003, ch. Interfacing the Narrative Experience.
- [13] FISK, R., AND GROVE, S. The Service Experience as Theater. *Advances in Consumer Research* 19 (1992), 455–461.
- [14] GOFFMAN, E. *The Presentation of Self in Everyday Life*. Doubleday: Garden City, 1959.
- [15] GRONBAEK, K., VESTERGAARD, P., AND ORBAEK, P. Towards Geo-Spatial Hypermedia: Concepts and Prototype Implementation. In *Proceedings of the 13th Conference on Hypertext and Hypermedia* (Maryland, USA, 2002).

- [16] HALASZ, F., AND SCHWARTZ, M. The Dexter Hypertext Reference Model. *Communications of the ACM* 37, 2 (February 1994), 30–39.
- [17] HARDMAN, L., BULTERMAN, D., AND ROSSUM, G. V. The Amsterdam Hypermedia Model: Adding Time and Context to the Dexter Model. *Communications of the ACM* 37, 2 (February 1994), 50–64.
- [18] HULL, R., CLAYTON, B., AND MELAMED, T. Rapid Authoring of Mediascapes. In *Proceedings of the UbiComp Conference* (2004).
- [19] JR., G. S. Time - The Next Source of Competitive Advantage. *Harvard Business Review* 86, 4 (1988), 41–51.
- [20] KRAY, C., KRUGER, A., AND ENDRES, C. Some Issues on Presentations in Intelligent Environments. In *First European Symposium on Ambient Intelligence (EUSAI)* (2003).
- [21] LAUREL, B. *Computers as Theatre*. Addison-Wesley, 1993.
- [22] MAES, P. How To Do The Right Thing. *Connection Science Journal* 1, 3 (1989), 291–321.
- [23] MAGERKO, B. A Proposal for an Interactive Drama Architecture. In *AAAI 2002 Spring Symposium Series: Artificial Intelligence and Interactive Entertainment* (2002).
- [24] MATEAS, M. An Oz-Centric Review of Interactive Drama and Believable Agents. Tech. rep., School of Computer Science, Carnegie Mellon University, Pittsburgh, 1999.
- [25] MCKEE, R. *Story: Substance, Structure, Style and The Principles of Screenwriting*. Regan Books, 1997.
- [26] MILLARD, D., AND E.A. Hyperdoc: An Adaptive Narrative System for Dynamic Multimedia Presentations. In *Proceedings of the 14th Conference on Hypertext and Hypermedia* (2003).
- [27] MURRAY, J. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. MIT Press, 1998.
- [28] PINE, J., AND GILLMORE, J. *The Experience Economy*. Harvard Business School Press, 1999.
- [29] RIVKIN, J., AND RYAN, M. *Literary Theory: An Anthology*. Blackwell Publishers, 1998.
- [30] ROMERO, L., AND CORREIA, N. HyperReal: A Hypermedia Model for Mixed Reality. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia* (Nottingham, UK, 2003), ACM.
- [31] SAWHNEY, N., BALCOM, D., AND SMITH, I. E. HyperCafe: Narrative and Aesthetic Properties of Hypervideo. In *UK Conference on Hypertext* (1996), pp. 1–10.
- [32] SCHECHNER, R. *Performance Studies: An Introduction*. Routledge: New York, 2002.
- [33] SCHULZE, G. *Die Erlebnisgesellschaft: Kultursoziologie der Gegenwart*. Campus, 1992.
- [34] SINCLAIR, P., MARTINEZ, K., MILLARD, D., AND WEAL, M. Links in the Palm of your Hand: Tangible Hypermedia using Augmented Reality. In *Proceedings of the 13th Conference on Hypertext and Hypermedia* (Maryland, USA, 2002).
- [35] SOHN, T., AND DEY, A. iCAP: Rapid Prototyping of Context-aware Applications. In *Proceedings of the CHI 2004 Conference on Human Factors in Computing Systems* (2004).
- [36] STOTTS, P., AND FURUTA, R. Petri-net-based hypertext: document structure with browsing semantics. *ACM Transactions on Information Systems* 7, 1 (1989), 3–29.

- [37] SYNCHRONIZED MULTIMEDIA INTEGRATION LANGUAGE (SMIL) W3C STANDARD. <http://www.w3.org/AudioVideo/>.
- [38] SZILAS, N. Interactive Drama on the Computer: Beyond Linear Narrative. In *AAAI 1999 Fall Symposium on Narrative Intelligence* (1999).
- [39] TOFFLER, A. *Future Shock*. Bantam Books, 1984.
- [40] WEAL, M., MICHAELIDES, D., TOMPSON, M., AND ROURE, D. D. The Ambient Wood Journals - Replaying the Experience . In *Proceedings of Hypertext '03* (Nottingham, UK, 2003).
- [41] WEISER, M. The Computer of the 21st Century. *Scientific American* (1991).

# Living with Hyper-reality

Leonardo Bonanni

MIT Media Laboratory,  
20 Ames Street, Cambridge, MA 02139 USA  
amerigo@media.mit.edu

**Abstract.** Hyper-reality describes distributed computing interfaces that weave existing environments with additional channels of sensory feedback to enhance everyday activities without confusing users. To be intuitive and non-intrusive these interfaces use illusionary pre-attentive content that is co-located with the objects and surfaces of a space and synchronous with a user's actions. Interfaces for an office, a laboratory, a kitchen and a public hallway are presented along with user studies suggesting that augmenting sensory feedback has the potential to simplify tasks and make them safer, while expanding the potential for interaction with everyday environments.

## 1 Introduction

In the early 1980s French social theorist Jean Baudrillard coined the term 'hyper-reality' to describe places that feel more real than the real world by blending an existing environment with simulated sensations [4]. A decade later Mark Weiser predicted a future of 'ubiquitous computing' in which intelligent machines distributed throughout the real world could sense and serve our everyday needs [34]. Since then many kinds of reality-based interaction have been proposed that seek to make digital interfaces more natural by distributing them through the objects and spaces of the real world. Virtual reality, augmented reality, tangible interfaces and ambient displays all propose means for adding new channels of digital information to the real world without overwhelming users. We are becoming accustomed to representation as a growing part of our lives. Much of our work and play occurs through computer interfaces. Projectors, television and computer monitors are becoming larger and the quality of simulated content more illusionary. It is becoming possible to merge the real world with simulations that enhance our actions and create new sensory experiences. At the same time, many of the tasks we perform on a daily basis are under-represented and lack feedback. The electric devices we rely on use simple sounds, lights and alphanumeric displays to inform us. Few of the appliances in our homes and offices can attract our attention when needed, leading to mistakes or just plain waste. In the modern kitchen, for example, mitigating the dirt and smell of cooking has correspondingly muted visual and tactile feedback, both useful to cooks. But enhancing feedback from everyday tasks is not only useful – it can be expansive, with the potential to make chores more attractive and to lend value to neglected resources. Reality-based interaction can have a real impact on the way we interpret the world around us and the actions we take. Hyperreality interfaces describes a merging of mundane tasks with intuitive, immersive interfaces that lend additional sensory experience. In the continuum traditionally used to characterize reality-based interaction, where virtual reality is at one extreme and reality at the other, hyper-reality can be seen as extending the spectrum of

how ‘real’ an experience feels by superimposing sensory simulation based on the existing environment (see Fig. 1). Informed by the status of people and tools in a space, hyper-reality interfaces magnify experience of everyday events in a manner that is intuitively understood at a sensory level. By mapping this intuitive sensory information directly to the objects and surfaces of everyday spaces, hyper-reality interfaces can provide greater confidence and control of basic tasks without interfering with others. Because much information is undesirable, these interfaces are designed based on the attention and comprehension of users to be intuitive and non-intrusive. Unlike Augmented Reality, hyper-reality is based solely on the real world, and only provides feedback rooted in the experience – not from external sources. For this reason, it can be considered more ‘real’ than ‘reality,’ especially when used in everyday spaces that are lacking sensory information. Overlaying a task as mundane as using a sink with sensory channels of light or sound can make users more conscious of their actions and have wide-ranging impact – from making the task more pleasant to reducing water waste or promoting better hygiene – without detracting from the task at hand. A number of hyper-reality interfaces for everyday environments have been designed over the past three years. Evaluations of these augmented sensory environments suggest that many mundane tasks could benefit from enhanced sensory feedback to become easier, more pleasurable and to motivate new behaviors.



**Fig. 1.** Hyperreality describes interfaces that enhance sensory perception of everyday experiences by layering additional channels of feedback

## 2 Related Work

Ubiquitous computing is making it possible to distribute computational functionality throughout the objects and spaces of the real world in ways that can be useful, non-distracting and expansive. These interfaces types range from totally artificial (virtual reality) to entirely physical (tangible media). They suggest that distributed interfaces can be easy to interpret so long as they are co-located, pre-attentive, and synchronous. In turn, illusionary interfaces can even expand on our sensory and perceptual cognition to promote new behaviors and effect new sensations.

Augmented Reality (AR) seeks to overlay the everyday world with useful information that is directly mapped to the objects and places it refers to. This is typically accomplished by wearing a head-mounted display that can draw text and graphics in the user’s field of view. Such task-intensive interfaces have been proposed for technicians repairing printers, astronauts on spacewalks or even surgeons during an operation [13,14]. Augmented Reality has the advantage of being ‘co-located’ or directly overlaid on a user’s focus of attention, so that it is easy to understand what a specific piece of information refers to. Information can also be projected on the surfaces of a space, making augmented reality suitable for public use by multiple users [29]. But projecting information in the form of text and graphics onto the real world drawbacks: it is cognitively intensive, requiring focused attention on the task, and cannot be

scaled. In the *Augmented Reality Kitchen*, graphics and text were projected on the countertops and appliances of a residential kitchen to assist and orient users through a recipe [8]. Users in a pilot study performed recipes more poorly when information was projected all around the environment as compared with following a hand-held recipe, in part because of the cognitive weight of distributed text and graphics and because users could not interpret sequential tasks simultaneously. The main benefit of AR is that it can place information where it can be intuitively understood: in the Media Lab's Counter-Intelligence lab, the temperature of meat in the oven is projected directly onto the food, in Microsoft's kitchen of the future, users can measure out how much flour is needed for a recipe by completely filling in a projected circle on the countertop [27].

Ambient displays offer a means to distribute information in everyday spaces without overwhelming users by communicating pre-attentively. *Pinwheels* are simple paper fans distributed through architectural space that spin to reflect traffic on the local computer server [21]. The *Stock Orb* is a glass bulb that glows a different color according to the performance of the stock market. Ambient displays are designed to be 'glance-able' or 'pre-attentive,' so that users can gather their information without needing to disrupt their principal task [1]. They are often placed at the periphery of vision, providing subtle information in the way that a window lets you remain conscious of the time of day and weather outside without requiring you to interrupt your work. Ambient displays are so subtle that they can also be difficult to notice, let alone comprehend: before an ambient display can be useful, a user must (1) know *that* it exists (2) know *what* it refers to and (3) know *how* to interpret the information [18]. For this reason ambient interfaces are often private, intended to appear decorative to all but their owners. But with intuitive, co-located content, ambient displays can enrich everyday consciousness without interfering with a primary task.

One solution to making interfaces intuitive and informative at the same time are Tangible User Interfaces (TUIs), where everyday objects are imbued with computational power [22]. One reason TUIs are so intuitive to learn is that they provide synchronous, co-located feedback directly mapped to the manipulation of commonplace objects. In *musicBottles*, empty glass bottles serve as 'containers' of digital information that can be physically un-corked to reveal a musical track [23]. In *Urp*, the wooden models of an architect are augmented with projected shadows allowing already useful objects to take on an added layer of information and experience [32]. Because synchronous audio-visual feedback occurs seamlessly with manipulation of the objects (bottles, models), even novice users can understand the relationship of their physical actions and the computer's augmentations. Applying such synchronous sensory feedback to many mundane actions can increase a user's confidence in the task they are performing, even when the feedback is mediated through novel channels of experience.

Reality-based computer interaction also suggests that computers can enrich our everyday lives by transforming mundane tasks into immersive experiences with expansive consequences. Virtual Reality (VR), in which are immersed in a totally artificial world through head-tracking display goggles, has been demonstrated to have significant effect on the perception of pain and fear. Studies show that burn patients can undergo wound treatments with substantially less pain when they are playing a VR game that takes place in a snowy world [17]. This case demonstrates the power of



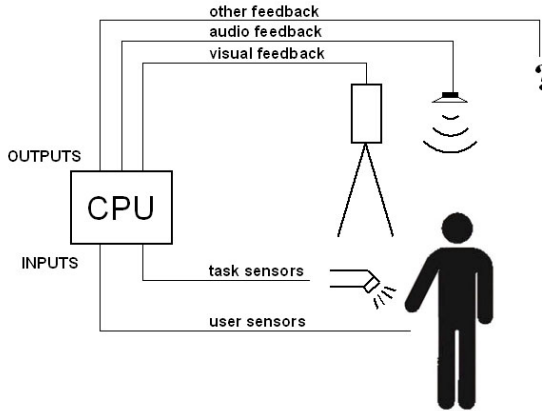
synesthesia or mapping one sense to another (soothing vision to tactile pain). Similar techniques have been demonstrated effective at reducing pain during chemotherapy, and even to help alleviate fear of flying, heights and spiders [30]. Reality-based interfaces can also be illusionary, with the potential to impact the way we perceive and interact on a daily basis. *InStink* uses computer-dispensed perfumes to enrich communication, so that you can smell dinner cooking at home from the office [26]. In *CounterActive*, a kitchen counter projected with recipe tutorials also had the ability to set a unique mood for each meal through sound and video [25]. *Nebula* is a bedroom ceiling projector that can more effectively soothe people to and from sleep by immersing them in a virtual sunset or dawn [28]. Reality-based interaction can also entice new behaviors by enriching everyday experiences, from guiding museum visits [35] to helping kids learn new recipes at home [25]. In one project, the immersive nature of AR is exploited to transform the daily walk to work into a real-world video game [11]. By making an everyday chore fun, such interfaces can encourage people to walk more often, cook at home or visit museums. Because ambient displays operate almost subconsciously, they have also been proposed to subtly motivate lifestyle changes leading to resource conservation, healthy habits and social contact [19,20]. One informal study by the makers of the *Stock Orb* notes that owners of the device check stock quotes less often on their computers, but they trade stocks more often, suggesting that ambient information can be persuasive by keeping users conscious of neglected information while not entirely focused on it [2]. Many persuasive techniques can be effective at motivating behavior change resulting in resource conservation and improved health and hygiene [12,15]. By taking advantage of reality-based interaction, interactive environments can motivate new behaviors without being distracting or confusing.

### 3 Design

Hyperreality interfaces seek to provide intuitive feedback that is not distracting and has the potential to motivate new behaviors enjoyably. The hypothesis is that overlaying sensory information on everyday actions can make people more conscious of processes in their environment as well as their own actions, from which they may choose to take on different behaviors. By carefully designing distributed interfaces to be co-located, synchronous, pre-attentive and illusionary they can more easily be accepted and provide positive benefits without distracting or confusing from the task at hand. This paper presents four hyper-realities that have been implemented and are being evaluated, beginning with a case study of a faucet. In every case, the system consists of an everyday experience measured by sensors and overlaid with digitally mediated sensory feedback (see Fig. 2).

#### 3.1 HeatSink

In the kitchen, users have come to depend on remote controls and indicators to know the temperature of food and water or the status of the stove. For example, how often do we scald ourselves at the faucet or wait arbitrarily for tap water to reach a desired



**Fig. 2.** Hyper-reality architecture: sensors detect the actions of a user and overlay the experience with additional channels of sensory feedback



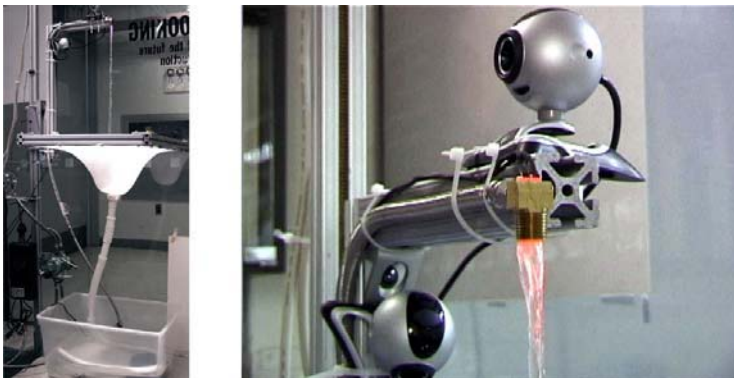
**Fig. 3.** *HeatSink* makes the temperature of tap water visible by projecting colored light into the stream

temperature? *HeatSink* is a simple solid-state circuit that projects colored light into the stream of tap water to indicate its temperature intuitively: red for hot fading to blue for cold. Taking a cue from projected augmented reality interfaces, the projection of colored light directly into the stream proves more successful than remote indicators like the control knob because the information is overlaid directly on the user’s focus of attention (see Fig. 3). During design, one iteration was considered that maps temperature to a full red-green-blue spectrum like the *Stock Orb*. The final choice of simple red and blue was based on the fact that people do not intuitively understand the temperature of ‘green’ water; their main concern is to determine whether the water is colder or hotter than their hands before touching it. By using various intensities of only two colors, *HeatSink* displays only the minimum essential information and does not inconvenience the task at hand or require prior knowledge.

In a study, 16 novice users aged 18-48 were asked to fill cups alternately with very hot and very cold water with or without the aid of the *HeatSink*. Observation and questionnaire answers reveal that over 90% (15/16,  $p < .05$ ) understood the colored light during their first use and were able to fill containers with hot or cold water without touching the stream, suggesting that this example of increased sensory feedback was able to motivate behavior change almost instantly. In addition to performing its function, the device has been seen by hundreds of visitors to the lab and their comments have helped understand why it is so effective. The synchronous illumination and illusory color make it immediately apparent what is going on. It is comforting to have feedback on something you are doing, even if you could find the information other ways. Finally, the lack of a need to touch the water makes itself evident when you see the light. This simple device prompted the development of more interfaces for the sink (*SmartSink*), the kitchen in general (*Cooking with the Elements*) and public spaces (*gurgle*). Since its first public presentation in 2004 [5], similar systems have been put in production by at least one faucet manufacturer [16].

### 3.2 SmartSink

Fresh water is one of our most important resources, yet the way drinking water is distributed effortlessly makes it likely to be ignored or wasted. Interaction with water can have serious consequences for health and hygiene as well – proper hand-washing is the most effective way to spread infection [10], yet many health care workers do not wash their hands as often as they should (less than half the time according to some studies [33]). I was part of an interdisciplinary team that sought to discover new interface possibilities for water with the *SmartSink*: a platform for experimentation consisting of a working, sensor-laden sink installed in a laboratory [3,6] (See Fig. 4). Digital video cameras mounted to the faucet are used to identify the kind of action being performed (washing hands, filling a container, washing fruits, etc...) through image recognition. Based on the action being performed, lights in the stream of water and a small speaker provide positive feedback to try and motivate water conservation,



**Fig. 4.** *SmartSink* showing the research platform (left) and a detail of the faucet with digital video cameras for image understanding (right)

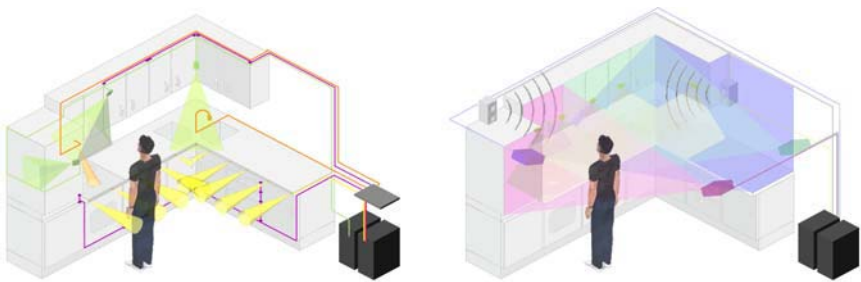
or proper hand-washing, depending on the context. Informal testing of this sink in the lab has revealed that light in the faucet and music or voice feedback are considered pleasant, not annoying or intrusive. By dedicating so much feedback to the sink and water itself, *SmartSink* makes people aware of various water-related issues, while the use of positive feedback alone as reward keeps the system from being irritating. For example, the sink congratulates good water conservation by playing a short piece of music, saying ‘thank you,’ or making a show of colored light in the water. Discussions with professionals from the health care and food service industry, however, reveal that more strict feedback would be desirable in critical ‘clean room’ application – one example has the door to an operating room on an electric lock, which can only be opened once satisfactory hand-washing is recorded. In a home or office environment, however, *SmartSink* could serve as a fun way to teach proper hygiene as well as a daily reminder about the preciousness of water.

### 3.3 Cooking with the Elements

I am a deaf individual, as in I cannot hear at all and use no assistive listening devices. I have always been a little annoyed that I do not get the full sensory information of all my household appliances. For example, I never know when the microwave is finished, or when the stove timer beeps. I could purchase signaling systems which involve pagers and whatnot that beep when a pre-recorded tone is detected, and such. But in my opinion, a sensory-rich environment is much better and more natural than wearing a pager. I have an overhead fluorescent light that likes to flicker just slightly when my washing machine kicks from the filling-up to the agitate phase, and flickers again (from just a coincidental power diversion) when the machine stops. I enjoy this sensory information much more than a pager buzz.

-Candace Myers

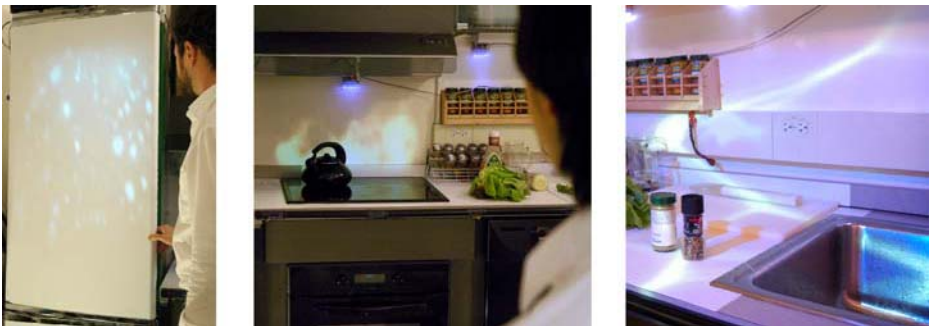
The modern kitchen is a technological marvel that combines the elements of fire, water, ice and earth in a compact hygienic space. The modern aesthetic combined with advances in hygienic materials have resulted in a space that can be surprisingly devoid of sensory experience, considering its function. *Cooking with the Elements* is a hyperreality that maps intuitive multimedia textures to the countertops of a conventional kitchen to enrich sensory feedback and inform tasks in the space [7]. Common problems such as knowing if the oven is hot or keeping the refrigerator door open too long can be intuitively annotated with dynamic audiovisual textures projected onto the surfaces of the appliances themselves.



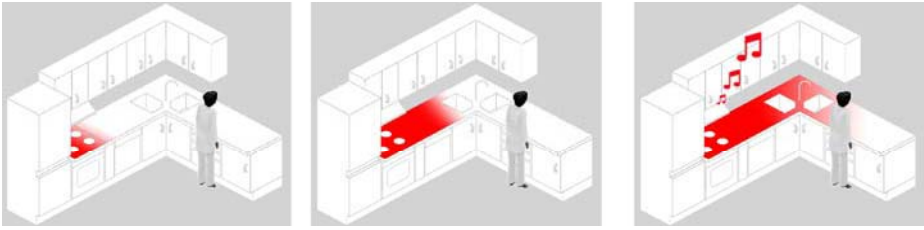
**Fig. 5.** *Cooking with the Elements* inputs (left) include proximity, temperature, motion and flow sensors while the outputs (right) consist of three tiled projections with stereo speakers

*Cooking with the Elements* is a networked ubiquitous computing environment in which sensors and effectors cover every part of a residential kitchen replica at the Media Lab. Proximity sensors situated along the countertop edge locate users while temperature and water sensors and micro-switches detect the status of the cabinets, countertops, sink, and appliances. Projectors seamlessly display on the countertops, appliances and cabinets while stereo speakers offer directional sound (see Fig. 5). As with *musicBottles*, tangible interaction with various appliances in the space releases synchronous, co-located sound and light projections that complement the experience intuitively. The projections are pre-attentive and *illusionistic*, using only readily-comprehended imagery in large projections at the periphery of vision. When someone opens the refrigerator, the sound of a cold wind plays and a projection of snow appears on the door. The snow accumulates to give an impression of how long the door stays open and how much energy is wasted. When the electric range is on or the stove reaches desired temperature, a dynamic fire is projected on the backsplash along with the crackling sound of a wood fire. If the sink is left running, a projected pool of water grows to cover the countertop while the sound of a bubbling creek fills the room (see Fig. 6).

Depending on where users are located, these displays grow or shrink to remain in the periphery of their attention and never to detract from their current task. The proximity sensors along the countertop edge provide a good idea of whether someone is using the kitchen or not. If they are, the display grows until it surrounds them on the countertop. An IR thermometer above the range can detect if the surface temperature of food is adequate. If the dish requires attention, a projection of fire grows to approach the user while remaining in the periphery of her vision. If this doesn't get noticed, the sound of crackling fire plays (See Fig. 7). The system also works if no one is in the room: in case a user forgets the water running or the stove on, the displays grow to fill the room so that anyone walking by the kitchen is immediately aware that something is wrong. Although the displayed textures only convey limited information (hot, cold, wet) they seek to do so in a completely intuitive manner that is always accessible and never distracting. *Cooking with the Elements* enriches the sensory nature of cooking and returns some of the feedback that was lost when kitchens became modern and hermetic.



**Fig. 6.** Cooking with the Elements overlays simulated feedback experience on the refrigerator (left), cook-top (middle), and sink (right)



**Fig. 7.** Illustration showing how ambient information grows to alert users in an intuitive, ambient manner: when sensors measure that the food on the cook-top needs attention, a projection grows to fill the countertop before a subtle acoustic reminder

User studies of *Cooking with the Elements* suggest that enhanced sensory feedback can make everyday tasks more easy to perform and keep certain types of mistakes from occurring. For example, the modern cook-top of the kitchen in the study is made of black glass, devoid of any indication of its temperature aside from some miniature light-emitting diodes. A study was conducted in which 16 novice users aged 18-48 were asked to write if they thought the cook-top was hot or cold while standing a few feet away. In a study, less than 20% (3/16,  $p < .05$ ) of people could determine that the cook-top was hot when standing a few feet away. With the projected fire, on the other hand, almost 90% (15/16,  $p < .05$ ) of people assumed the cook-top was hot. This simple example reveals that modern appliances can under-inform users as to their status. In a more subjective study, users were asked to retrieve an item from the freezer with or without the projection and sound of a blizzard and write what they felt. Nearly half of users (7/16) reported feeling cold or rushed to close the door when the simulated storm was playing, suggesting that such feedback could motivate people to change their behavior by amplifying sensory experience. The sense of touch was overlaid with the two additional channels of sight and sound to make an experience effectively feel colder, and the urgency of closing the refrigerator door greater. Finally, the interfaces were evaluated in terms of hedonics, or how desirable they are. *HeatSink* and the cook-top were preferred, probably because their function was more immediately perceptible to users, whereas the refrigerator was slightly less desirable, probably because it made some users actually feel cold.

### 3.4 Gurgle

Do you drink enough water? Many times dehydration can set in without any noticeable symptoms [31]. Fellow Media Lab researcher Ernesto Arroyo and I conceived an interface for motivating people hurrying down a public hallway to stop and take a drink from a water fountain. *gurgle* is an interactive installation for the water fountains on the MIT campus, where they are often relegated to dingy nooks. When someone walks by the drinking fountain, a shimmering blue light entices them to approach. If they take a drink, they are rewarded with a sound-and-light show: a watery reflection fills the entire space along with the sound of a babbling brook. Proximity sensors have been installed on site for several months, so that usage data can be obtained from the water fountain with and without *gurgle*. The system randomizes feedback, so that the importance of light and sound can be better understood. The

architecture is modular, so that many water fountains can be augmented at low cost. It consists of sensors to detect the actions of people and the status of the machine (in this case the drinking fountain), feedback through audio and video projection, and a built-in microcontroller that directs feedback while serving to log user preference. Implemented over the long term, these interfaces will have to provide feedback on a varying schedule, trying to maintain the novelty of the experience so that it continues to be effective without becoming oppressive. Currently one version of *gurgle* is installed at a public fountain on the MIT campus (see Fig. 8) and another is slated to be installed in the lobby of a mixed-use high-rise.



**Fig. 8.** *gurgle* augments the refreshing experience of drinking from a hallway water fountain to entice people to stay hydrated

## 4 Conclusion

Many experiences and environments do not provide enough feedback to be fully valued and understood. The spaces we inhabit can be overlaid with sounds, images and other sensations through ubiquitous sensors and displays. Information can be distributed throughout everyday spaces if it is carefully designed to be easily understood and non-intrusive. Co-located projection of illusionary information operating synchronously with a user's actions helps informative environments remain intuitive. Immersive, intuitive feedback can have transformative effect on the physiological perception of a space. Neglected spaces and tasks can become immersive and enriching, and new environments can be made easier to approach. By simply magnifying the feedback that occurs during everyday activities, users can be made more conscious of positive or negative behaviors. The hyper-realities described in this paper only enhance sensations related to architectural spaces, but future interfaces could also enhance our perception of people and social situations as well – in the way that a monitor speaker informs a musician about how her sound is perceived by the audience. Ubiquitous computing can make the world around us richer and more beautiful by expanding our sensory perception of the everyday.

## Acknowledgements

I wish to thank Ernesto Arroyo, Michael Barrett, Chia-Hsun Lee, Subodh Paudel, Sam Sarcia, and Jon Wetzel for their help implementing the interfaces in this paper, and Candace Myers for her inspiring candid opinion.

## References

- [1] Ambient Devices: [www.ambientdevices.com](http://www.ambientdevices.com)
- [2] Ambient Devices CEO David Rose
- [3] Arroyo, E., Bonanni, L., and Selker, T. *Waterbot: Exploring Feedback and Persuasive Techniques at the Sink*. Long paper in proceedings of Computer Human Interaction (CHI) 2005, Portland, OR.
- [4] Baudrillard, Jean. *Simulacra and Simulation*. Tr. Sheila Faria Glaser. Ann Arbor: University of Michigan. 1994. Originally published in French by Editions Galilee, 1981.
- [5] Bonanni, L., Lee, C.H. *The Kitchen as a Graphical User Interface*, in SIGGRAPH 2004 Electronic Art and Animation Catalog, 109-111.
- [6] Bonanni, L. Arroyo, E. Lee, C.H., Selker T. *Ambient intelligence: the next generation of user centeredness: Exploring feedback and persuasive techniques at the sink*, ACM interactions, July 2005 Volume 12 Issue 4.
- [7] Bonanni, L., Lee, C.H., Selker, T. *Cooking with the Elements: Intuitive Immersive Interfaces for Augmented Reality Environments*. In Proc. INTERACT 2005, Rome, Italy.
- [8] Bonanni, L., Lee, C.H., Selker, T. *Counter Intelligence: Augmented Reality Kitchen*. Long paper in Extended Abstracts of Computer Human Interaction (CHI) 2005, Portland, OR.
- [9] Bonanni, L., Lee, C.H., and Selker, T. *Attention-Based Design of Augmented Reality Interfaces*. In Proc. CHI '05, Portland OR.
- [10] Centers for Disease Control [[www.cdc.gov](http://www.cdc.gov)]
- [11] Cheok, A.D., Goh, K.H., Liu, W., Farbiz, F., Fond, S.W., Teo, S.L., Li, Y., Yang, X. *Human Pacman: a mobile, wide-area entertainment system based on physical, social, and ubiquitous computing*. Personal Ubiquitous Computing (2004) 8:71-81.
- [12] Cialdini, R. *The science of persuasion* Scientific American, 2001, 76-81.
- [13] Feiner, Steven K. *Augmented Reality: a New Way of Seeing*. Scientific American April 2002.
- [14] Feiner, S., MacIntyre, B., and Seligmann, D. (1993). *Knowledge-based augmented reality*. Communications of the ACM, 36(7):52{62}.
- [15] Fogg, B.J. *Persuasive Technology: Using Computers to Change What we Think and Do*. Morgan Kaufmann, 2002.
- [16] Hansa faucets: <http://www.hansa.de>
- [17] Hoffman, H.G. *Virtual-Reality Therapy*, Scientific American August 2004.
- [18] Holmquist, L. E. *Evaluating the Comprehension of Ambient Displays*, in Proc. CHI 2004 pp. 1545.
- [19] S.S. Intille, C.K., R. Farzanfar, and W. Bakr, *Just-in-Time Technology to Encourage Incremental, Dietary Behavior Change*. in AMIA 2003 Symposium, (2003).
- [20] Intille S.S. A new research challenge: persuasive technology to motivate healthy aging. *Transactions on Information Technology in Biomedicine*, 8 (3).2004.
- [21] Ishii, H., Ren, S., and Frei, P. *Pinwheels: Visualizing Information Flow in an Architectural Space*, in Proc. CHI '01, pp.111-2.



- [22] Ishii, H. & Ullmer, B., Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In pRoc. CHI '97, pp. 234-41.
- [23] Ishii, H. Fletcher, R., Lee, J., Choo, S., Berzowska, J., Wisneski, C., Cano, C., Hernandez, A., Bulthaup, C., *musicBottles*, in Proc. SIGGRAPH '99, pp. 174.
- [24] Jacob, R. *Reality-Based Interaction: A New Framework for Understanding the Next Generation of Human-Computer Interfaces*, white paper at <http://www.eecs.tufts.edu/~jacob/theory/>
- [25] Ju, W. et. al. (2001). *Counteractive: An Interactive Cookbook for the Kitchen Counter*, in Extended Abstracts CHI 2001, 269-70.
- [26] Kaye, J. N. (2001) *Symbolic Olfactory Display*. Master's Thesis, MIT Media Lab, 2001.
- [27] Microsoft Kitchen of the Future as seen in the Food Network's documentary 'Kitchens of the Future,' 2003.
- [28] Philips Nebula:  
<http://www.design.philips.com/about/design/section-13534/index.html>
- [29] Podlaseck, M., Pinhanez, C., Alvarado, N., Chan, M., Dejesus, E., *On Interfaces Projected onto Real-World Objects*, in Proc. CHI 2003.
- [30] Rauterberg, M. *Positive Effects of VR Technology on Human Behavior*. In Proc. ICAT '04 International Conference on Artificial Reality and Telexistence, pp. 85-88.
- [31] Saltmarsh, M. *Thirst: or, Why do People Drink?* In Nutrition Bulletin, 26, 2001, pp. 53-58.
- [32] Underkoffler, J., Ishii, H. *Urp: a Luminous-Tangible Workbench for Urban Planning and Design*. In Proc. CHI '99, pp. 386-93.
- [33] University of California at San Francisco – Stanford University Evidence-based Practice Center. *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*. Agency for Healthcare Research and Quality, Contract No. 290-97-0013. [<http://www.ahrq.gov/clinic/ptsafety/>]
- [34] Weiser, M. "The Computer for the Twenty-First Century," *Scientific American*, pp. 94-10, September 1991
- [35] Woods, E., Billinghamurst, M., Aldridge, G., Garrie, B. *Augmenting the Science Center and Museum Experience*. Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia, 2004, pp.230-6.

# Ambient Pre-Communication

## A Study of Voice Volume Control Method on Telecommunication

Atsunobu Kimura<sup>1</sup>, Yoshihiro Shimada<sup>1</sup>, and Minoru Kobayashi<sup>2</sup>

NTT Cyber Solutions Laboratories  
NTT Corporation

1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847 Japan

<sup>1</sup> {kimura.atsumobu, shimada.yoshihiro}@lab.ntt.co.jp  
<sup>2</sup> minoru@acm.org

**Abstract.** We present Ambient Pre-Communication for always-on telecommunication systems. It allows the Caller to feel more comfortable in initiating communication, as comfortable as in the real world. To start communication in the real world, determining the appropriate voice volume is essential. However, existing tele-communication systems fail to support the Caller in controlling voice volume because the systems don't provide feedback of the Caller's voice at the Receiver's site. Our proposal allows existing tele-communication systems to provide visual feedback. The key idea is to superimpose a ripple-peak-meter, which represents the sound field at the Receiver's site, on the Receiver's video signal. The Caller can control his/her voice volume using the ripple-peak-meter. We describe a prototype system and experiments in which we observed the Caller's response to the ripple-peak-meter display. The proposed method allows the Caller to easily grasp and control the volume of his/her voice at the Receiver's site and encourages him/her in commencing tele-communication.

## 1 Introduction

Always-on tele-communication is more easily realized these days by the spread of broadband networks. Always-on tele-communication is a service that allows users to conduct a conversation using the video and audio signals provided by the network to other sites. This service is assisted by the enhancement of small communication devices such as displays, cameras, loudspeakers, and microphones. These communication devices well support the user's communications needs. As a result, environments like VideoWindow [4], MediaSpace [1],[8],[11] and other always-on tele-communication systems [5],[13],[14] are being adopted.

Always-on tele-communication systems offer significant advantages in that the Receiver's situations can be readily observed, but many users hesitate to employ these systems. One of the worries is that their voices will be reproduced inappropriately at the Receiver's sites when the Caller wants to call the Receiver; the result is that the Caller feels stress in initiating communication.

For example, your friend, who has a young baby, is at her home and you are at your home. You may not be so eager to call her at night because of concerns that your ring will wake the baby. This is a very natural concern and must be alleviated if these systems are to be used more widely. In another example, you want to ask your friend to go to a restaurant on the weekend, which is not urgent task, but the friend appears to be busy at the remote site and is not paying attention to your call. In this example, you hesitate to interrupt your friend and thus would like to speak softly to confirm his/her ability to enter into a conversation.

The above problems and other similar problems are easily resolved in face-to-face communication. Unfortunately, they are not well tackled by existing tele-communication systems.

This paper introduces the concept of Ambient Pre-Communication; it allows the Caller to confirm the actual volume of his/her voice as regenerated at the remote site. To realize this concept, we use a ripple-peak-meter to show the Caller's voice as it is recreated at the remote site. We discuss an implemented system, and present the results of an initial study that demonstrate the effectiveness of the approach; we show that visual feedback makes the Caller more comfortable in initiating tele-communication.

## 2 Goals of Ambient Pre-Communication

When a person tries to start a conversation, he/she moves through the following stages; discerns if the Receiver is available (idle etc.) for communication or not (busy etc.) [Discern stage]; tries to address the Receiver at the voice volume desired by him/her [Try stage]; discerns the Receiver's reaction, or lack of reaction, to the his/her call [React stage]. We define these stages as Pre-Communication. It starts with the decision to communicate and ends with the commencement of communication.

Face-to-face communication allows the Caller to easily realize all three Pre-Communication stages, unlike existing tele-communication systems. One essential difference is the ability to control voice volume in the Try stage; existing tele-communication systems isolate the Caller from his/her Receiver and make it impossible for the Caller to assess his/her voice volume as reproduced at the Receiver's site. We believe that the lack of this ability in the Try stage is a critical concern and explains why so many people hesitate to employ tele-communication systems.

To resolve this problem, we propose the concept of Ambient Pre-Communication. It is defined as an environment with ambient devices linked via communication networks to provide Pre-Communication like face-to-face communication for the Callers using tele-communication systems. Ambient Pre-Communication significantly reduces the stress placed on the Caller of a tele-communication system. Our aim is to ensure that any Ambient Pre-Communication system can

support the Try stage with the ability to control Caller's voice volume as reproduced at the Receiver's site.

Here is one story that illuminates the effectiveness of Ambient Pre-Communication.

*A father is at his office and his wife and child are at home.  
When the child should be going to sleep,  
the father is trying to call his wife.*

*He uses a system that offers Ambient Pre-Communication.  
He can see his wife tucking the child into bed.  
The proposed scheme allows him to lower his voice,  
so that it doesn't disturb the child*

This example shows the need for the Ambient Pre-Communication system to provide an adequate understanding the audio field and control of the voice volume at the Receiver's site in the Try stage.

### 3 Related Work

The research focus of this paper is a system that encourages a Caller to start communication with a Receiver at a remote site by allowing the Caller to control his/her voice volume at the remote site. Our approach is to visualize the voice volume in a feedback channel; the resulting system encourages the Caller by providing a face-to-face like communication environment in daily-life situations.

The naive solution is direct auditory feedback of the voice volume as reproduced at the Receiver's site to the Caller. This is not practical because the Caller finds it very difficult to speak while listening to a slightly delayed version of his/her own words. Our solution is visual feedback to achieve Ambient Pre-Communication.

In the 1990s, tele-communication systems with video audio channels became a hot research topic. Systems such as VideoWindow[4] were proposed but their adoption rates were quite low. Subsequent research attempted to enhance the attractiveness of the video channels by providing different communication channels such as "awareness" [3]. The concept of awareness includes gaze awareness [10] and proximity awareness [18]. Other channels mentioned involved haptic technology [15] and olfactory effects [16]. While previous research is useful in terms of maintaining interest in an existing communication session, no serious effort was made to encourage the establishment of the session, Pre-Communication.

Communication in virtual reality environments is a rather special case since current systems require both parties to enter highly specialized rooms with large amounts of hardware. Obviously Pre-Communication is not considered by these systems since it is assumed that their desire to use the system is strong enough.

They do not encourage spontaneous communication or communication in many real-life situations [2],[7],[12],[17].

InterActor [19] is an interesting approach which has the aim of encouraging both parties to continue the communication session by overcoming the weakness of audio-only connections. The system analyzes the Receiver's speech to extract some form of "emotion" and uses the information to drive the Receiver's avatar to generate more satisfying communication. Once again, this system does not explicitly address Pre-Communication.

Some research has tackled interfaces for the hearing impaired. For safety outside the house, a system was proposed that captured ambient sounds and displayed a visual representation of them to the user [9]. Two display forms were mentioned. One was a simple ripple shaped display to show the locations of sound sources and sound volumes. The other, a spectrum display, allowed the user to make some judgment as to the nature of the source. In addition, in the art field many different types of sound visualization have been proposed but none were created from the point of view of encouraging the start of communication. ([6] is one of them).

## 4 Proposed Method

The visual effect is an indication of the Receiver's sound field which is inferred from the audio volume at the Receiver's loudspeaker. The visual effect is made to overlay the remote Receiver's video; its metaphor is a ripple. Fig.1 shows a basic implementation with a Caller using the system. The ripple-peak-meter, which comes from the center of the Receiver's loudspeaker (right bottom of Fig.2), expresses the sound field of the Caller's voice. To realize this method, we need to consider how to display the feedback as well how to infer the Receiver's sound field. This paper examines the first step.



**Fig. 1.** Implementation of the ripple-peak-meter



Fig. 2. Display of the ripple-peak-meter at Caller's site

## 5 Experiment

### 5.1 Experiment's Design

We implemented the ripple-peak-meter, which allows the Caller to grasp how well the Receiver and the general public can hear the Caller, in an experimental

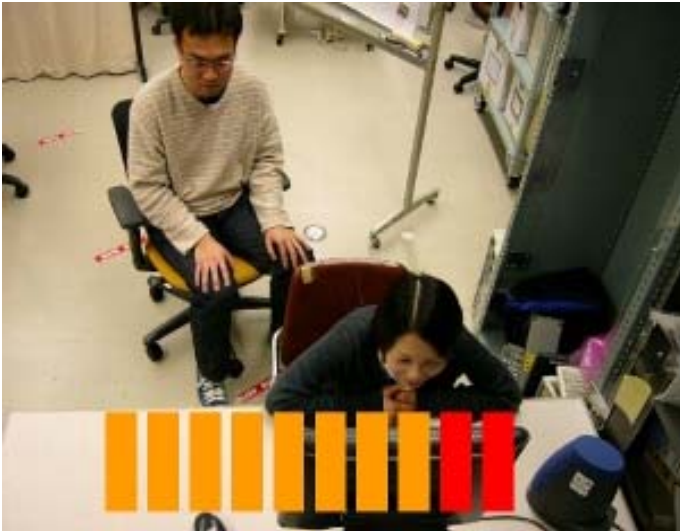
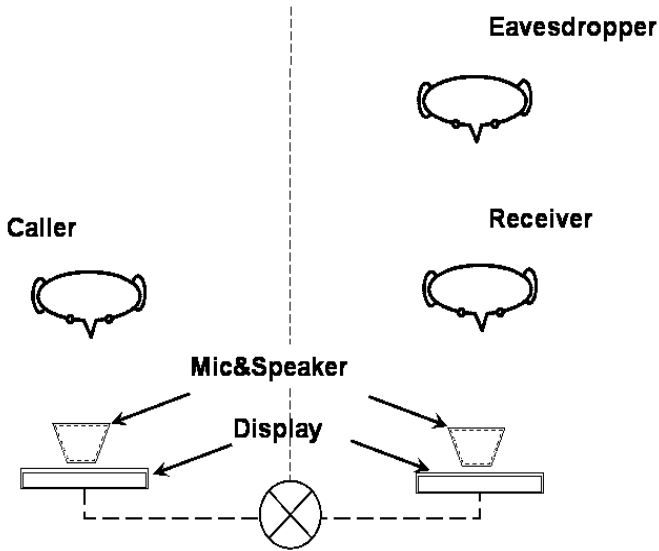


Fig. 3. Example of input-peak-meter



**Fig. 4.** System structure at Caller's site(right-side) and Receiver's site(left-side)

system. This system was tested to confirm its effectiveness in controlling voice volume and its impact on Ambient Pre-Communication.

A video-audio tele-communication system was implemented that used a ripple-peak-meter, an input-peak-meter (shown in Fig.3), or no peak-meter. The input-peak-meter is a regular peak-meter that uses bar graphs to directly indicate the volume level of the Caller's microphone input (see the bottom of Fig.3).

Two or three people participated in each experiment. The Caller attempted to speak to the Receiver or to initiate communication with the Receiver, without allowing the Eavesdropper to hear the Caller's voice. Figures 4,5 and 6 show the system structure and scenes of the experiments at the Caller's site and the



**Fig. 5.** A scene of experiment: Caller's site



**Fig. 6.** A scene of experiment: Receiver's site

Receiver's site. The Caller could see the Receiver's video overlaid with the ripple-peak-meter or the input-peak-meter. The Caller's site was isolated from the site shared by the Receiver and the Eavesdropper. The Receiver and the Eavesdropper occupied the same site with the latter further from the loudspeaker.

The ripple-peak-meter and the input-peak-meter were implemented using Flash Player and a Flash Communication Server. For simplicity, the ripple-peak-meter was fed from the Caller's microphone input and thus only approximated the loudspeaker output at the Receiver's site. One way of implementing the proposed system in an actual application is to install an additional microphone near the loudspeaker at the Receiver's site. Accurate calibration is possible by installing more microphones at the Receiver's site.

## 5.2 Experiment 1 (Voice Volume Control)

**Purpose and Conditions.** We tested whether the Caller could control voice volume as desired with the ripple-peak-meter. The task was to whisper words written on a list without allowing the Eavesdropper to hear the Caller's words, only the Receiver could hear them.

Input volume of Caller's voice and the number of words captured by the Receiver and the Eavesdropper were measured. Two subjects acted as Caller in this experiment.

The settings of the input microphone and Receiver's loudspeakers were fixed to overboost the Caller's voice. This required the Caller to speak softly to achieve reasonable sound levels at the Receiver's site. The settings were not informed to the subjects. By using a very small voice, the Caller could whisper to the Receiver.

In this paper, subject A or B were tested as Callers with the input-peak-meter or the ripple-peak-meter so 4 combinations were tested: ExpA-input, ExpA-ripple, ExpB-input, and ExpB-ripple.

**Procedure.** The video audio tele-communication system, the input-peak-meter and the ripple-peak-meter were explained to the subjects before the experiment.

The Caller was required to read as many Japanese words as possible from a list given beforehand in one minute to just the Receiver. The given list (see the example in Table 1) was extracted from "List of Japanese words used for distinction level test for defective Receiver" and consists of very common Japanese words of about 4 letters.

The Receiver and Eavesdropper were forbidden to respond to the Caller and were instructed to just transcribe the words. The reason for this was to prevent the Caller from using their responses to control his/her voice volume so that the Caller had to control the voice volume by using just the ripple-peak-meter. For each utterance of the Caller, the Receiver and Eavesdropper filled in one row of a data table; the entry was either the word heard or <blank> if the word was not understandable.

**Results (Input Volume on Microphone).** Figures 7, 8, 9 and 10 show input volume over time for ExpA-input, ExpA-ripple, ExpB-input, and ExpB-ripple, respectively.

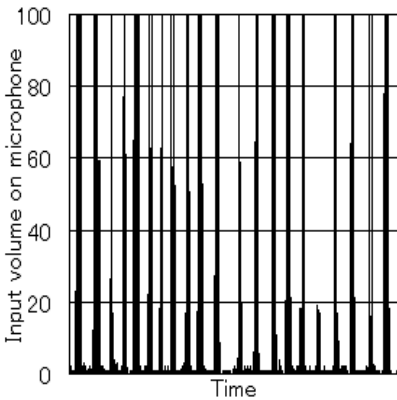


**Table 1.** Some of the words used

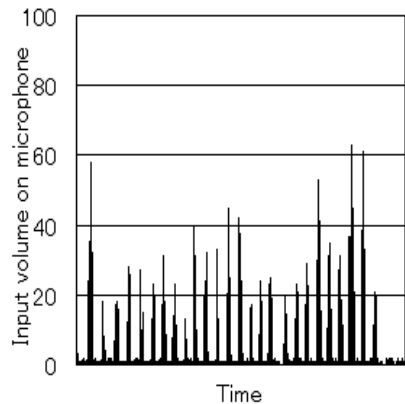
|    | Japanese words | English meaning |
|----|----------------|-----------------|
| 1  | Oyamoto        | One's home      |
| 2  | Keisatsu       | Police          |
| 3  | Jitsubutsu     | Original        |
| 4  | Soramimi       | Imagination     |
| 5  | Tsunagari      | Connection      |
| 6  | Ninnniku       | Garlic          |
| 7  | Hikidashi      | Drawer          |
| 8  | Yamakaji       | Forest fire     |
| 9  | Ryakudatsu     | Loot            |
| 10 | Mijinnko       | Water flea      |
| 12 | :              | :               |
| 13 | :              | :               |

Figures 7 and 9 show that, relative to the predefined "best" input volume (30 on the vertical axis), the Callers failed to control their voice volume appropriately when the input-peak-meter was used. This is because the Caller's voice was overboosted, and the input-peak-meter made it difficult for the Caller to realize voice loudness at the Receiver's site.

Figures 8 and 10 show that the ripple-peak-meter allowed both Callers to control voice volume properly.



**Fig. 7.** Result of ExpA-input



**Fig. 8.** Result of ExpA-ripple

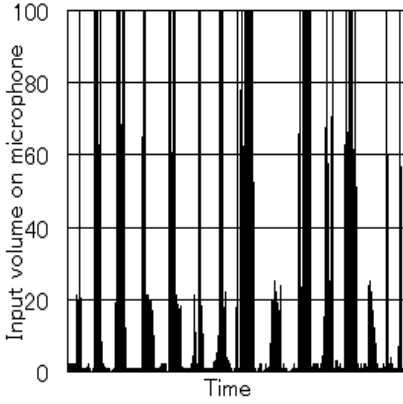


Fig. 9. Result of ExpB-input

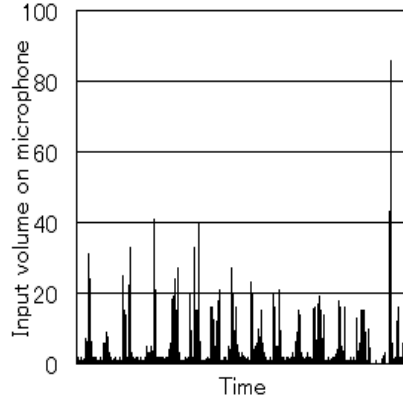


Fig. 10. Result of ExpB-ripple

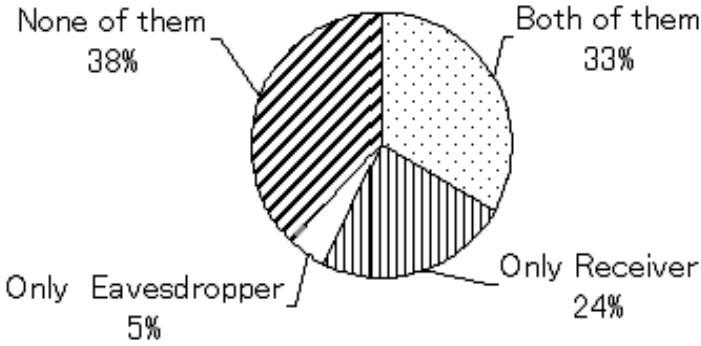


Fig. 11. Result of ExpA-input

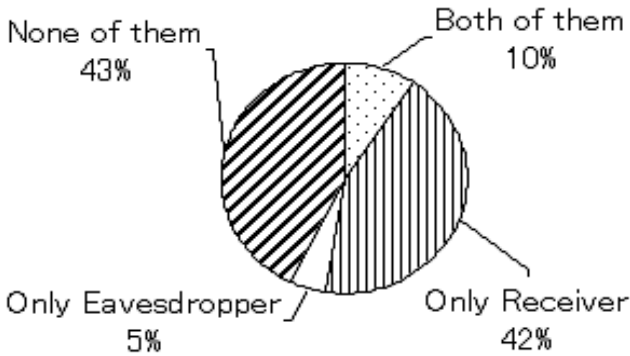


Fig. 12. Result of ExpA-ripple

**Results (Number of Words Captured).** Figures 11 and 12 show the ratio of the number of words captured to the number of all words the Caller spoke for ExpA-input and ExpA-ripple, respectively.

Figures 11 and 12 show that the ripple-peak-meter was much better in suppressing the leakage of words to the Eavesdropper. While zero leakage was the goal, the results indicate that the Eavesdropper had somewhat better hearing ability than the Receiver, so this goal was basically unattainable. In both situations, ExpA-input and ExpA-ripple, the rate of "None of them" was rather high (38% and 43%). The reason is that the task set, speaking isolated words, made it difficult for the Caller to speak each word at the same volume.

### 5.3 Experiment 2 (Practical Situation)

**Purpose and Conditions.** To observe the effectiveness of the ripple-peak-meter in a task completion situation and to examine the psychological impact on the Caller or Caller's behavior, we examined the proposed system in a message passing task, which involves Pre-Communication. The system with ripple-peak-meter was compared to a conventional system without ripple-peak-meter.

The experiment tested three Caller tasks and three Receiver states. Task-urgent directed the Receiver to send a message within 30 seconds. For Task-private, the Caller attempted to pass a message to the Receiver without leaking information to the Eavesdropper. The instruction given for Task-casual was to invite the Receiver to a social event while discerning the Receiver's state. The Receiver was instructed to either hold a conversation with someone else (State-conversation), to concentrate on a PC-game (State-game), or to be idle (State-idle).

The system was set to yield three responses. Volume-straight yielded basically the same sound level. Volume-boost increased the sound level at the Receiver's site; Volume-suppress did the reverse. In the experiment, the Caller could use either the tele-communication system or a telephone to pass the message; that is, the telephone replaced the microphone input as the sound capture device and the Receiver heard the Caller via a regular telephone handset.

The number of calls until the Receiver reacted and Caller's words (said in the experiment) and comments (collected in an interview after the experiment) and video-observation data were gathered.

**Procedure.** Subjects completed a tutorial in which they were introduced to the basic functionality of the video-audio communication system and the ripple-peak-meter. After the tutorial, they experienced both systems at each remote site.

In the experiment, when the Caller clicked the start button, a 5 sec. count-down was displayed and after 5 sec. the message and some instructions were displayed. At the Receiver's site, the same 5 sec. count-down was shown together with the instruction to the Receiver indicating the state to be adopted. When the counter reached 0, the Receiver was waiting in the appropriate state, and the Caller then tried to pass the message indicated. Regardless of the Receiver's state, for Task-private and Task-casual, each trial was terminated after 2 minutes. Task-urgent trials were terminated in 30 seconds.

**Table 2.** Conditions for Experiment 2

| Receiver's State | Volume setting | Caller's task |         |        |
|------------------|----------------|---------------|---------|--------|
|                  |                | casual        | private | urgent |
| conversation     | straight       | o             |         |        |
| game             | straight       | o             |         |        |
| idle             | straight       | o             | o       |        |
| idle             | boost          |               |         | o      |
| idle             | suppress       |               |         | o      |

Each pair of subjects was instructed to perform 12 trials. Each subject performed 6 trials of the conditions shown in Table 2, with and without the ripple-peak-meter. The ordering of the conditions was balanced among the subjects.

**Results (Ripple-peak-meter Ensures Rapid Response).** The Caller's words and subsequent interview comments demonstrate the value of the ripple-peak-meter. The words and comments were analyzed together with the video-captured data.

Figures 13, 14 and 15 show the numbers of Caller's calls until the Receiver reacted.

Figures 13 and 14 show that even in very relaxed communication situations [conditions: Task-casual, Volume-straight], the ripple-peak-meter improved the response speed of the Receiver.

A tougher communication situation is shown in Figure 15 since the sound level was decreased and the message was urgent. In this tough environment, the absence of the ripple-peak-meter greatly increased the number of calls needed, and thus the delay. Feedback from the two Callers indicated the value of the ripple-peak-meter in the Task-casual condition.

A typical comment after experiment without ripple-peak-meter is given below.

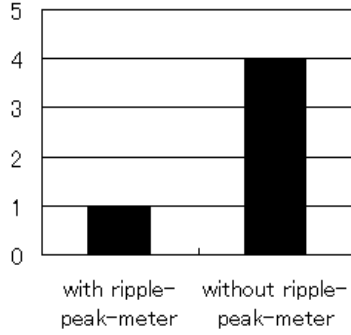
*"Without the ripple-peak-meter it took longer to control voice volume properly, for example, I asked the Receiver about the loudness of my voice again and again."*

**Results (Caller Reassurance).** The comments of the subjects yielded two main findings. First, the Pre-Communication step is quite stressful and is a source of tension. Second, the ripple-peak-meter was seen as a strong support tool. One subject commented before the experiment:

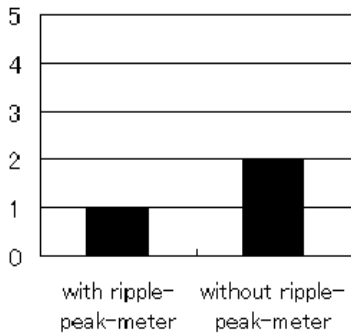
*"I don't need the ripple-peak-meter because it makes it difficult to see the Receiver's video."*

After the experiment, the same subject commented:

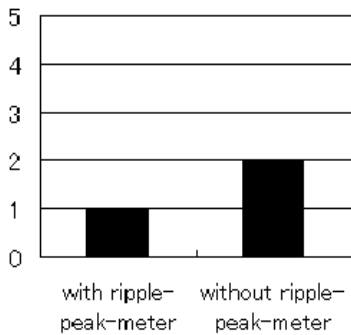
*"By looking at the ripple-peak-meter, I could keep my voice volume at the appropriate level."*



**Fig. 13.** Numbers of calls until Receiver reacted [conditions: Task-casual, Volume-straight, State-game]



**Fig. 14.** Numbers of calls until Receiver reacted [conditions: Task-urgent, Volume-suppress, State-idle]



**Fig. 15.** Numbers of calls until Receiver reacted [conditions: Task-casual, Volume-straight, State-game]

During the experiments we noted that when the Receiver was allowed to respond in Experiment 2, the Caller sometimes expressed confusion between the display of the ripple-peak-meter and response of the Receiver. This is entirely natural since once communication has been established between people, they tend to emphasize direct feedback to control the communication session. The ripple-peak-meter is most effective for Pre-Communication since the Caller has no direct feedback.

## 6 Discussion

The results of our experiments showed that the ripple-peak-meter offered the following effects; the Caller could grasp and control his/her voice volume to the Receiver, the ripple-peak-meter reduced the numbers of calls in certain situations, and the ripple-peak-meter encouraged the Caller to start communication on the tele-communication system.

These benefits result from the characteristics of our ripple-peak-meter. One is positional relation in the system display between the ripple-peak-meter and the superimposed video image at the Receiver's site. It makes it easy for the Caller to grasp the Caller's voice at the Receiver's site. Even through the ripple-peak-meter is displayed on a 2 dimensional surface (current version), the system provides enough spatial clues since the camera is fixed. In this paper, the camera was set to provide a view midway between a bird's-eye view and a full-face view. The ripple-peak-meter uses the wave metaphor, which has the same properties as sound. A louder source yields ripples that spread more widely. This is a simple response and allows the Caller to control his voice intuitively.

We intend to improve the ripple-peak-meter in several ways. The experiments described in this paper restricted the usage positions of the subjects. Given our goal of encouraging the adoption of these systems in daily-life, we need to develop robust sensing technology that can accurately grasp the sound field at the remote site. It is important to develop a simple and effective calibration method.

To provide useful feedback for communication, we need to investigate various ways of representing sound or voice. Many various visual effects other than the ripple-peak-meter could be used to provide useful feedback. One idea is to devise a way of tagging people at the remote site with visual representations of the sound level as they perceive it. Moreover, auditory feedback, haptic feedback or olfactory feedback might be possible. Of course, multi-modal feedback would provide useful feedback for communication, especially for the physically challenged or the people engaged in parallel tasks. One of the most important requirements in designing any feedback system is that it allow the Caller to grasp the range of his/her voice at a certain volume, for example minimum audible volume. We will clarify requirements for feedback systems through iterated cycles of designing and testing.

The systems and experiments described in this paper were designed to focus on the Caller because he/she is key person starting the conversation. However, the Receiver is also important. In order not to annoy the Receiver or to disturb

the Receiver's task, we will identify better design requirements, for example the Receiver can understand the Caller's concern for the Receiver or the Receiver can check the Caller's disturbance, through experiments to observe the Receiver's behavior and the changes in behavior that occur during a conversation.

## 7 Conclusion

Ambient Pre-Communication was proposed for always-on tele-communication systems; it overlays a ripple-peak-meter representation of the sound field at the remote Receiver's site on the Receiver's video signal as feedback for the Caller. A prototype system was constructed and experiments were conducted to assess the effectiveness of the ripple-peak-meter and the subjects' behaviors.

The results of experiment 1 confirmed that the ripple-peak-meter allowed the Caller to easily grasp and control the level of his/her voice at the Receiver's site.

Experiment 2 yield subjective Caller's comments and measured data; both indicated that the ripple-peak-meter encouraged the Caller to initiate tele-communication sessions because the Caller felt reassured that the Receiver would respond to the Caller's well modulated voice. Tests in which the number of calls needed to attract the Receiver's attention was measured showed that the ripple-peak-meter and Ambient Pre-Communication approach allowed the Caller to control the voice volume when conducting limited tasks and yielded more rapid response by the called party.

We need further experiments in real situations to confirm objectively the validity of our approach.

## Acknowledgment

The authors would like to thank our colleagues in the Human Interaction Project for their participation in our experiments, and Dr. Katsuhiko Ogawa, Dr. Ken-ichiro Shimokura and Mr. Masayuki Ihara for their helpful comments on this paper.

## References

- [1] Bly, S., Harrison, S. and Irwin, S.: Media spaces: Bringing people together in a video, audio and computing environment. *Communication of the ACM*, Vol.36. (1993) 28–45
- [2] Cohen, M., and Koizumi, N.: Exocentric Control of Audio Imaging in Binaural Tele-Communication. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, Vol.E75-A(2). (1992) 164–170
- [3] Dourish, P. and Bellotti, V.: Awareness and coordination in shared workspaces. *Proc. of the 1992 ACM conference on Computer-supported cooperative work*, (1992) 107–114
- [4] Fish, R. S., Kraut, R. E., Chalfonete, B.: The VideoWindow System In Informal Communications. *Proc. of the 1990 ACM conference on Computer-supported cooperative work*, (1990) 1–11

- [5] Fish, R. S., Kraut, R. E., Root, R. W.: Evaluating Video as a Technology for Informal Communication. Proc. of the 1992 ACM Conference on Computer Human Interaction, (1992) 37–48
- [6] Golan, L. and Zachary, L.: In-Situ Speech Visualization in Real-Time Interactive Installation and Performance. Proc. of The 3rd International Symposium on Non-Photorealistic Animation and Rendering, (2004)
- [7] Greenhalgh, C., and Benford, S.: MASSIVE: a collaborative virtual environment for teleconferencing. ACM Trans. on Computer-Human Interaction, Vol.2. Issue.3. (1995) 239–261
- [8] Heath, C.C., and Luff, P.K.: Media Space and Communicative Asymmetries: Preliminary Observations of Video-Mediated Interaction. Proc. of the 1992 ACM conference on Human Computer Interaction, (1992) 315–346
- [9] Ho-Ching, W. H., Mankoff, j. and Landay, A. J.: Peripheral and ambient displays: Can you see what i hear?: the design and evaluation of a peripheral sound display for the deaf. Proc. of the SIGCHI conference on Human factors in computing systems, Vol.5. No.1 (2003) 161–168
- [10] Ishii, H., Kobayashi, M., G.: Integration of interpersonal space and shared workspace: ClearBoard design and experiments. ACM Transactions on Information Systems, Vol. 11. (1993) 349–375
- [11] MacKay, W. E.: Media Spaces: Environments for Informal Multimedia Interaction. Proceedings of the 1992 ACM conference on Computer-supported corroborative work, (1999) 55–82
- [12] Rodenstein, R., and Donath, J. S.: Talking in circles: designing a spatially-grounded audio conferencing environment. Proc. of the SIGCHI conference on Human factors in computing systems table of contents, (2000) 81–88
- [13] Root, W. R.: Design of a multi-media vehicle for social browsing. Proc. of the 1988 ACM conference on Computer-supported cooperative work, (1988) 25–38
- [14] Roussel, N.: Experiences in the design of the well, a group communication device for teleconviviality. Proc. of the tenth ACM international conference on Multimedia, (2002)146–152
- [15] Scott, B. and Andrew, D.: inTouch: A Medium for Haptic Interpersonal Communication. Proc. of the CHI 97 Conference on Human Factors in Computing systems, (1997) 22–27
- [16] Siiio, I. and Mima, N.: Meeting Pot: Coffee Aroma Transmitter. ACM UbiComp 2001: International Conference on Ubiquitous Computing, (2001)
- [17] Singer, A., Hindus, D., Stifelman, L., and White, S.: Tangible progress: less is more in Some wire audio spaces. Proc. of the SIGCHI conference on Human factors in computing systems, (1999) 104–111
- [18] Tang, J. and Rua, M.: Montage: Providing teleproximity for distributed groups. In Proc. of the 1994 ACM Conference on Computer Human Interaction, (1994) 37–43
- [19] Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R.: InterActor: Speech-Driven Embodied Interactive Actor. International Journal of Human-Computer Interaction, Vol.17. No.1. (2004) 43–60



# AmbientBrowser: Web Browser in Everyday Life

Satoshi Nakamura<sup>1</sup>, Mitsuru Minakuchi<sup>1</sup>, and Katsumi Tanaka<sup>1, 2</sup>

<sup>1</sup> National Institute of Information and Communications, Japan,  
3-5, Hikaridai, Seika-cho Soraku-gun, Kyoto, 619-0289, Japan  
{gon, mmina}@nict.go.jp  
<http://www2.nict.go.jp/jt/a133/gon/index.html>

<sup>2</sup> Graduate School of Informatics, Kyoto University  
Yoshida Honmachi, Sakyo, Kyoto, 606-8501, Japan  
ktanaka@i.kyoto-u.ac.jp  
[http://www.dl.kuis.kyoto-u.ac.jp/~tanaka/index\\_j.html](http://www.dl.kuis.kyoto-u.ac.jp/~tanaka/index_j.html)

**Abstract.** Recently, due to the remarkable advancement of technology, the ubiquitous computing environment is becoming a reality. People can directly obtain information anytime from ubiquitous computer. However, conventional computing style with a keyboard and a mouse is not suitable for everyday use. We proposed and developed a Web browser called the *AmbientBrowser* system that supports people in their daily acquisition of knowledge. It continuously searches Web pages using both system-defined and user-defined keywords, and displays sensors detect users' and environmental conditions and control the system's behavior such as knowledge selection or a style of presentation. Thus, the user can encounter a wide variety of knowledge without active operations. It monitors the context of the environment, such as lighting conditions and temperature. In addition, it displays Web pages incrementally in proportion to the context. This paper describes the implementation of the *AmbientBrowser* system and discusses its effects.

## 1 Introduction

In the words of Bertrand Russell, “*There is much pleasure to be gained from useless knowledge.*” Undoubtedly, people enjoy acquiring knowledge. They read books or watch TV to find out about trivia, even though these facts are of no practical use in everyday life. In Japan, for instance, one of the most popular TV programs called “*Trivial Fountain*” provides a variety of trivial knowledge to viewing audiences.

People may also compete in quizzes or parade their knowledge in conversation. This desire for knowledge can be considered to be related to the needs for esteem and being described in Maslow's hierarchy of needs [7]. Thus, we believe that intellectual stimulation is essential for human well being and the acquisition of new knowledge is a source of pleasure.

Many people acquire a great deal of knowledge from books, magazines, newspapers, posters, televisions, and computer displays. These media vehicles can be divided into *movable* and *immovable media*.

- *Movable media* (e.g., books, magazines, and newspapers)
- *Immovable media* (e.g., posters, advertising displays, and guide plates)

When people begin to read *movable media*, they have an explicit motivation to do so. When people start to read/watch *immovable media*, on the other hand, they have a lower motivation to do so than reading movable media vehicle, in many cases. We can say that the former is *active* and the latter is *passive browsing*. We believe that *passive browsing* is suitable for acquiring daily knowledge because, in our daily lives, we either relax or get involved with numerous activities. In addition, the variety of knowledge acquired by *passive browsing (immovable media)* may be greater than that gained by *active browsing (movable media)* because it targets all the information within the user's range of vision and the style of acquisition is heuristic. However, people cannot acquire much more knowledge by browsing a second time because such *immovable media* are static.

The ubiquitous computing environment is becoming a reality [20] due to the recent remarkable advances with computer technology. People can directly obtain information from ubiquitous computers through ubiquitous displays. There are many electronic billboards that provide news information dynamically to those waiting for it. They can acquire information naturally and pass away idle hours profitably. However, almost all electronic billboards only provide specific headline news, weather news, and advertisements. It is also difficult for people to read/browse content-rich information from them because almost all these billboards only have a single line display to represent text and this disappears almost immediately. Moreover, they cannot interact with electric billboards to control the speed at which information is represented.

Various methods for peripheral presentation of information have been proposed in related work. However, in some of these studies, involving what are called ambient displays, information is expressed in abstract form, e.g., using trembling, rotation speed, or figures in an image [1] [15]. Although these methods are less intrusive, they provide little information. Some researchers have proposed presenting detailed information on peripheral displays [2] [8], but these have not considered any enrichment of content. For example, how to cook delicious stews or who is a well known football player because almost all the information that is provided on displays is about news and weather.

Consequently, the main objective of our work was to achieve a mechanism that could provide a huge variety of richness content in the form of peripheral information to people in their daily lives. We assumed an environment that was surrounded by ubiquitous displays where these were located in places, such as kitchens, bathrooms, bedrooms, studies, offices, and streets. Computers that are connected to the Internet control these displays and we use them to provide peripheral information. We also use Web pages as peripheral information because there is an innumerable number of valuable Web pages in the World Wide Web (WWW).

We propose the *AmbientBrowser* system in this paper and explain its design and implementation. Then, we discuss its effectiveness and introduce some mechanisms to improve our system.

## 2 Ambient Browser

We assumed that the distance between people and ubiquitous displays would range from one to three meters in their daily lives. This means that they are not too close to

ubiquitous displays such as those on desktop computers and not too far from ubiquitous displays such as those on buildings. People read/browse Web pages on ubiquitous displays in the *AmbientBrowser* system environment [16].

The main methods of currently accessing Web pages are via linking and search engines. Both are active browsing methods, i.e., the user actively accesses information with a specific aim. We thus have to provide an automatic information retrieval mechanism. The Web browser, on the other hand, cannot render the entire Web page, which has been constructed with long passages of text and several images, on the window without a scroll bar. If users want to read/browse all the Web page, they have to scroll down/up to change the area being displayed. However, people in these environments do not like to use keyboards and mouse. Figure 1 shows nonsense situations where it is almost impossible to use keyboards and mice use in daily life.



**Fig. 1.** Nonsense situations

We had to consider the following in this work:

- How to select the target Web page and
- How to browse the target Web page.

*ViaVoice* [3] and *SmartVoice* [9] are methods of audio input for computers. They monitor the user's voice with a microphone and translate his/her voice into various text/commands based on the analyzed voice data of an individual. Users can control computers by only using their voice. This allows users to input keywords for searches and to navigate links. However, it forces them to first read out a large quantity of text to train their computers. In addition, this method is not resistant to noise. In addition, audio-input methods are not suitable for use in public spaces.

Igarashi and Hughes proposed using nonverbal features in speech to control interactive applications [4]. For example, they proposed continuous voice and rate-based parameter control through pitch. Users could change the area being displayed by using a simple voice command and tonguing. However, it was also not resistant to noise. In addition, it was extremely restricted for use in certain situations. For example, using this method in a public space is awkward, socially.

Some work has been done on gesture inputs. Gesture inputs with a camera and so on [16] allow users to operate applications without them having to wearing control devices. Users can control computers through gestures. However, they have to do gestures within the field of view of the camera. *Ubi-Finger* [17] [18] is a method of gesture input where users wear devices that allow them to do various operations. However, it is not convenient to wear these devices on a daily basis. There is also the

disadvantage of accuracy of recognition of gestures. Users who are relaxing would not want to use these devices every day.

The system discussed in this work selects the target Web pages automatically. In addition, it allows users to change the speed at which they are reading/browsing. The following subsections provide the details on how the target Web page is selected and browsed.

## 2.1 How to Select the Target Web Page

We believe that ubiquitous displays should have their own distinctive characteristics and roles because people can understand and acquire knowledge naturally if information that is provided by ubiquitous displays is related to their location or space. For example, someone at an aquarium can easily understand the ecology and the characteristics of fish if the ubiquitous display there has information about them. Someone in a grocery shop can easily determine how nutritious foods are and the cooking required if the ubiquitous display in a grocery shop has information about foods, drinks, and cooking. In the same way, ubiquitous displays at entrances should have information about doors, shoes/boots, and weather for users going on outings. Ubiquitous displays on trains should have information about trains, time tables, travel, and local events.

Ubiquitous displays may also have to reflect people's preferences for what information is to be provided. For example, if someone near a ubiquitous display wants information about sports news, then football, basketball, or baseball news may be provided. When people near a ubiquitous display want information about food, then data on restaurants, grocery stores, or cooking information may be provided.

To fulfill these requirements, we used various keywords and a Web search engine, such as Google search<sup>1</sup> or MSN search<sup>2</sup>. Our system selected target Web pages through keywords and these Web search engines. The keywords were divided into two types:

- Keywords for ubiquitous displays (Each ubiquitous display had several roles related to spatiality. For example, a ubiquitous display in the kitchen had "kitchen", "food", "cooking", "water", and "drink" as keywords. A ubiquitous display at an entrance had "shoe", "boot", and "weather" as keywords. In addition to these, locality keywords such as "Tokyo" and "Osaka" could have been used. These keywords are set by the owner.)
- Keywords for people's preference (Individuals have interests in various things. If someone has an interest in football, pasta, or Italian wine, these words are set as keywords. If someone has an interest in topical news, i.e., cake or scuba diving, these words are set as keywords. These keywords are set by the user.)

When the *AmbientBrowser* system selects a target Web page, it first randomly decides the number of keywords for the Web search by using a preset range of values. It then

---

<sup>1</sup> <http://www.google.com/>

<sup>2</sup> <http://www.msn.com/>

randomly selects various keywords from the ubiquitous display's reference keywords and people's preference keywords. It posts the selected keywords to the Web search engine to create a URL list of search results, and randomly selects the target Web page from the URL list. If the target Web page was displayed within the past few hours, it reselects the target Web page.

The target Web page (ambient information) is selected through these processes.

## 2.2 How to Browse the Target Web Page

Typically, browsing the Web requires active reading. In other words, the reader has to focus on following the text. He or she also has to manipulate the browser to update the pages being displayed by using the scroll bar or clicking on various links. This active browsing style, however, may not be suitable for everyday use. Most of the various activities we do daily do not involve sitting in front of a computer. It is difficult to operate a computer to browse Web pages using traditional interactive user interfaces that require direct manipulation, such as graphical user interfaces. Additionally, the user may find these operations tiresome when he/she is relaxing.

The conversion of Web pages into a TV-program-like format (text to audio, camera action, and avatar motion) [5] may be useful while relaxing in daily life. A TV program has time-based content, i.e., its representation proceeds automatically as time elapses. Users can thus watch content with no operations required by them. They can also watch time-based content actively by adjusting the current timing for playing time-based content. However, TV-program-like representation is not suitable for a ubiquitous display environment because the sound information that is provided by this representation may become ambient noise. In addition, it cannot provide information continuously.

We took the following into account so that we could provide ambient information to people in any place naturally:

- Passive browsing,
- Easily viewable representation, and
- Natural interaction.

Hence, by extending the concept of a time-based representation of content, we introduced gradual Web rendering as the core mechanism for the *AmbientBrowser* system based on an abstract parameter that could be connected to various statuses or inputs (temperature, time, brightness, and energy consumption). Our system enables users to control abstract parameters. The system changes the rendering speed depending on the abstract parameters. User can change the rendering speed by controlling an abstract parameter (see Fig. 2).

Gradual Web rendering first selects the target Web page and acquires the selected Web page from the WWW. It then divides this Web page into many parts and serializes them (see Fig. 3).

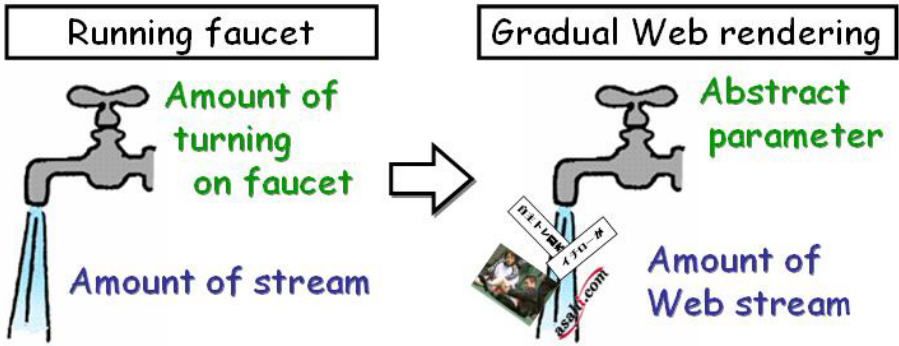


Fig. 2. Gradual Web rendering. User can change rendering speed by changing abstract parameter.

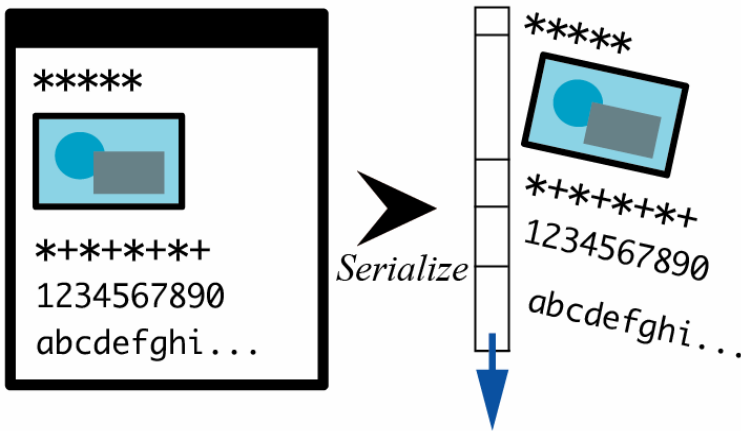


Fig. 3. Serialization of Web page

After this, it monitors the values for status or input from sensors or input devices, and converts them into part of the Web page. Finally, it renders part of the Web page after conversion (see Fig. 4). If the displayed fragment exceeds the size of the display area, it automatically scrolls down the display area. People can change their reading (browsing, rendering) speed by adjusting the parameter given by the sensor's output with this system. In addition, it can leave information for people to access for a certain time.

If we use brightness as an input parameter, the greater the brightness, the faster the browser displays the Web pages, and vice versa. Users can adjust the rendering speed by turning on a light at the sensor, or by shading the sensor with their hand. When the brightness is under a specific threshold, the *AmbientBrowser* system reduces the value of the parameters. Thus, the user can fast-forward or slow-forward the displayed page through a simple interaction.

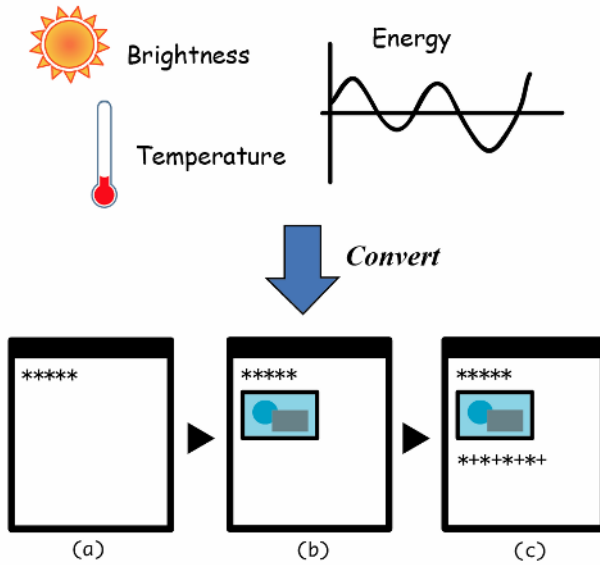


Fig. 4. Gradual Web rendering

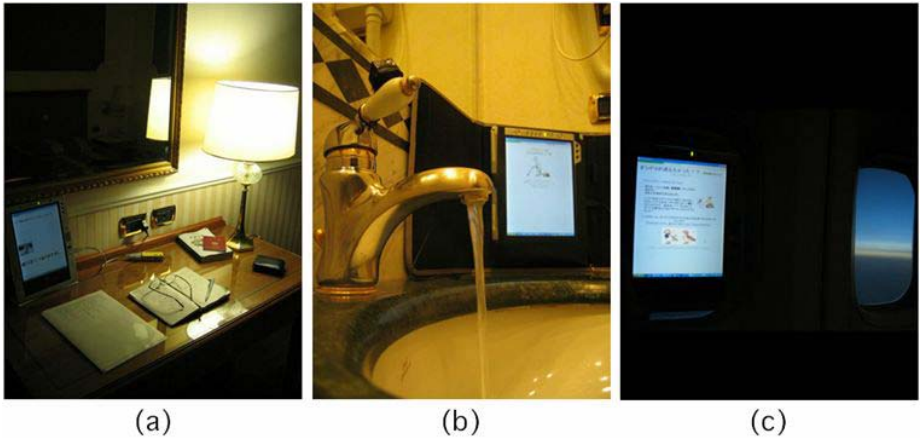


Fig. 5. (a) *AmbientBrowser* system uses brightness monitored by a sensor as parameter in study room. (b) It uses running faucet as parameter in bathroom. (c) When embedded display in window was used, it used distance from user to display as parameter.

Figure 5 shows examples of the *AmbientBrowser* system in use. In the study room in (a), the owner uses brightness monitored by a sensor as the parameter. He/she can view Web pages intermittently provided by the system when working. When he/she becomes interested in a Web page, he/she can read it slowly by locating his/her hand on the sensor. The *AmbientBrowser* system is in a bathroom in (b), where the owner has set a running faucet for the input parameter. There could be a situation where there are three faucet handles in the bathroom. The first faucet handle is for cold water, the second is for hot water, and

the third is for the Web. People can easily adjust the rendering speed. In addition, the input device fits the space. In (c), the *AmbientBrowser* system is embedded in the window and uses the distance from the user to the display as the parameter. If people approach it, it accelerates the rendering speed. At the same time, if people walk away from it, it slows the rendering speed. These are natural interactions.

### 3 Implementation

Figure 6 outlines the core mechanism for Web selection with the *AmbientBrowser* system. We used Radio Frequency Identification (RFID) to detect users in this implementation. We also used a keyword list that was created by the user from the WWW to acquire his/her preferences. We could set common keywords in the keyword list and the URL for the RSS (RDF site summary), such as the news and weblog sites.

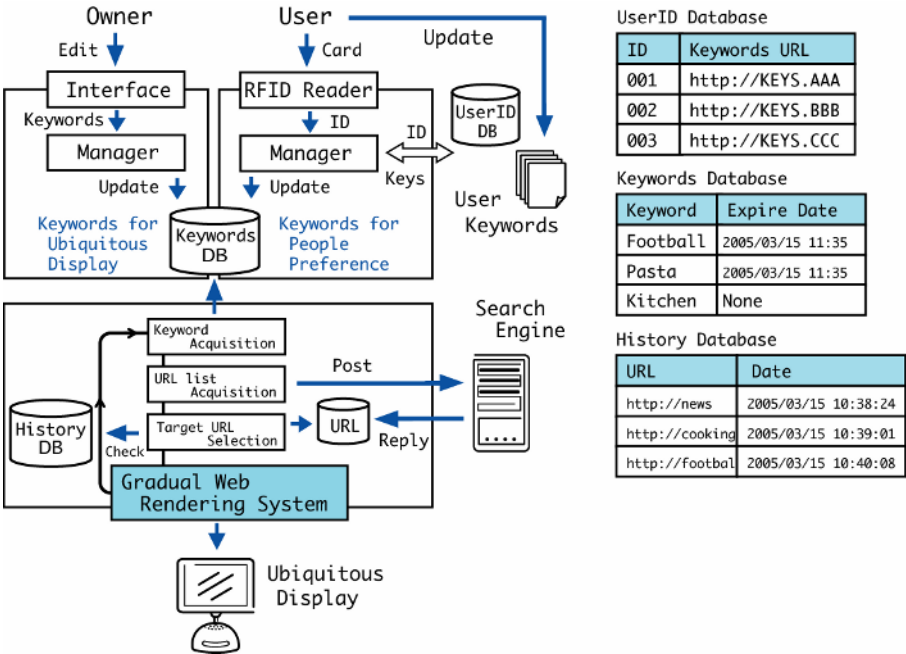


Fig. 6. Core Mechanism of Web selection for *AmbientBrowser* system

Each ubiquitous display's owner first sets keywords related to characteristics. For example, the owner of a ubiquitous display in a kitchen sets keywords like kitchen, cooking, food, or the RSS of a cook's weblog. In addition, he/she creates a keyword list of his/her preferences, updates it with the WWW, and registers his/her card ID with the User ID database with the URL of his/her keyword list.

When the user touches the RFID reader with his/her RFID card, our system detects the ID and obtains his or her preference keywords from a Web page. The manager of the user's preference keyword database checks the keywords and deletes ones that have expired.



When our system has selected the target Web page, the keyword acquisition module first randomly decides the number of keywords for the Web search by using the preset range of values, and randomly selects several keywords from the keyword database. The URL list acquisition module then posts the selected keywords to the Web search engine to create the URL list of search results. If the URL of the RSS was selected as a keyword, it creates the URL list from the RSS. The target URL selection module randomly selects the target Web page from the URL list. If the target Web page was displayed within the past few hours, it reselects the target Web page. After these processes, it sends the target URL to the gradual Web rendering system and adds the target URL to the history database.

We can see a photograph of the RFID card and RFID reader in Figure 7 (a). Figure 7 (b) is a keyword list of the user's preferences on the WWW.

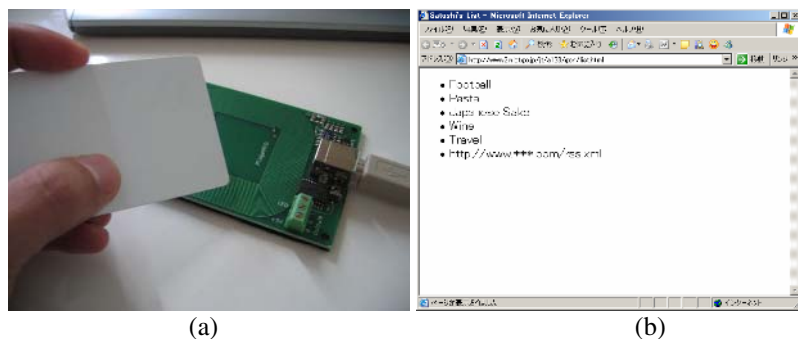


Fig. 7. (a) User registering preferences. (b) Keyword list of his/her preferences.

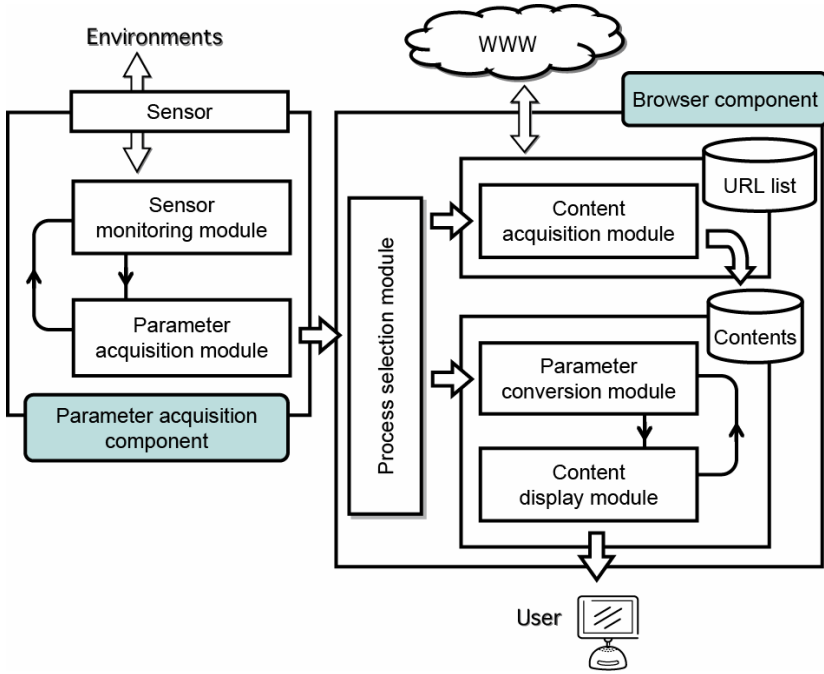
Figure 8 outlines the core mechanism for gradual Web rendering for the *Ambient-Browser* system. The system acquires parameters from the sensor and the Web page from the WWW. It formalizes the Web page and renders parts incrementally related to input parameters. The core system consists of the parameter acquisition component and the browser component.

- **Parameter Acquisition Component:**

The sensor-monitoring module constantly monitors the output such as wave, static value and so on by the sensor. The module sends the output data to the module that acquires the parameter. The parameter-acquiring module calculates the value of the parameter from the received outputs and sends the value to the browser component.

- **Browser Component:**

The process-selection module controls the browser processing according to whether it is in a waiting or processing mode. The waiting mode means that the browser has displayed an entire Web page and is preparing for the next page. The processing mode means that the browser is displaying a Web page. In the waiting mode, this module commands the content-acquisition module to select the next Web page and download it from the World Wide Web. In the processing mode, the module commands the energy-conversion module to render the Web page loaded by the content-acquisition module.



**Fig. 8.** Core mechanism for gradual Web rendering for *AmbientBrowser* system

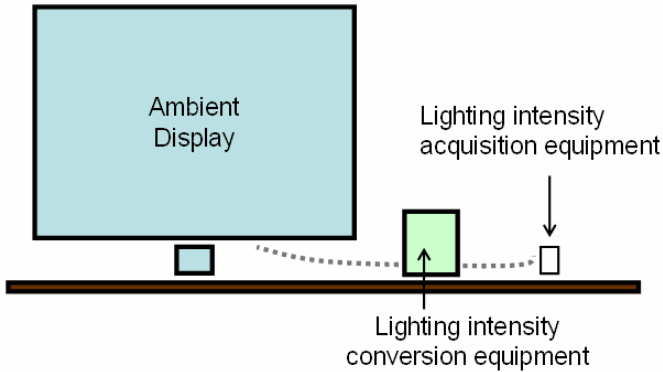
The content-acquisition module selects the next Web page from a URL list. This method eliminates the need for a separate operation for link navigation. After the module finishes downloading the target Web page, it stores the page, and deletes the page from the list. When there are no pages remaining on the list, it regards the program as finished and prompts the user to choose another one.

The parameter-conversion module generates a partial page to be displayed by cutting the original page at the point corresponding to the input parameter. This process conceptually converts the value of parameter into content. The conversion rate is preset in bytes per value. The user can adjust the speed at which Web content is rendered through the conversion rate.

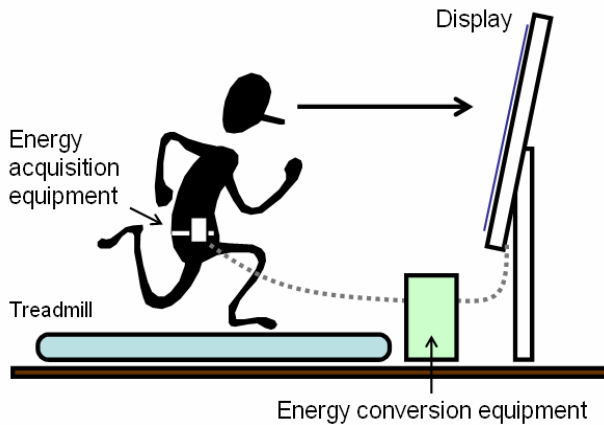
The content-display module renders the partial page generated by the parameter-conversion module. In addition, it scrolls down the display area automatically if the displayed fragment exceeds the size of the display area.

Figure 9 outlines the scheme for the *AmbientBrowser* system [11] using lighting intensity (brightness). The system consists of lighting-intensity-acquisition and lighting-intensity-conversion equipment and a ubiquitous display.

Figure 10 outlines the scheme for the *AmbientBrowser* system using a user's energy consumption (known as *EnergyBrowser* system [12] [13]). The system consists of energy-acquisition and energy-conversion equipment and a ubiquitous display.



**Fig. 9.** Scheme for *AmbientBrowser* system using lighting intensity



**Fig. 10.** Scheme for *AmbientBrowser* system using user's energy consumption (known as *EnergyBrowser* system)

Figure 11 has examples of gradual Web rendering processes. In (a), the system has rendered a blank page. In (b), the system has rendered the title of the Web page. In the (c), the system has rendered the main text. In (d), (e) and (f), the system has rendered various images and texts incrementally.

We used the VersaPro VY11F/GL-R produced by NEC Corp. as the ubiquitous display and computer, and the WeatherDuck produced by IT WatchDogs Inc.<sup>3</sup> for the brightness, temperature, humidity, volume of sound and air flow sensor (Fig. 12 (a)). Figure 12 (b) shows one example of the *AmbientBrowser* system using brightness, temperature, humidity, and air flow.

<sup>3</sup> <http://www.itwatchdogs.com/>



Fig. 11. Gradual-rendering process by *AmbientBrowser* system

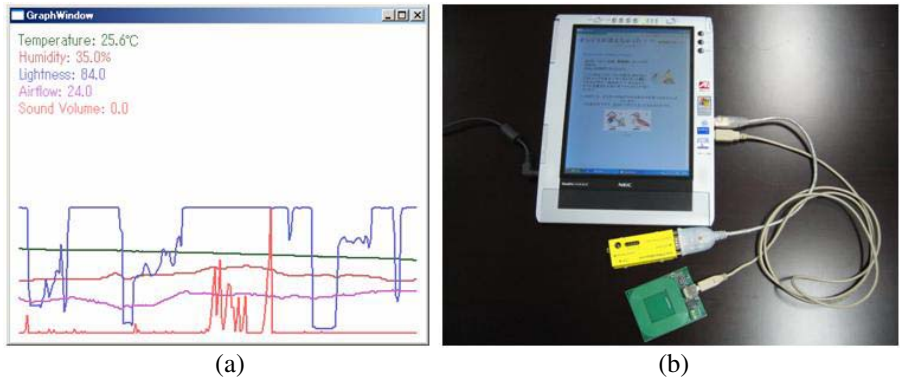


Fig. 12. (a) Sensor output. (b) Construction of *AmbientBrowser* system.

We also used the MDP-A3U9S produced by NEC Tokin Corp<sup>4</sup>. as the motion sensor (see Fig. 13 (a)) which has a ceramic gyro, acceleration sensor and terrestrial magnetism sensor for the *AmbientBrowser* system using energy consumption. In this system, we use the value of the terrestrial magnetism of it (see Fig. 13 (b)). Each wave corresponds to movement on each axis. The system calculates the energy consumed from the received waves. The waves have two parameters that reflect the consumed energy: frequency and intensity. For walking or jogging, the frequency reflects the number of steps taken by the user, and the intensity reflects the impact of each step. For ease of implementation, our system counts the steps according to changes in the waves and detects three levels of intensity. The total consumed energy is calculated by accumulating the intensity of each step.

<sup>4</sup> <http://www.nec-tokin.com/english/>

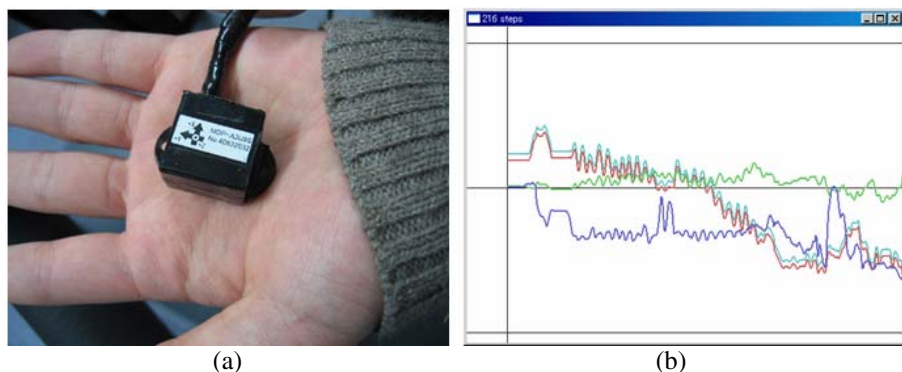


Fig. 13. Sensor and output for *AmbientBrowser* system using energy consumption

In addition, we used the RFID reader produced by Phidgets Inc.<sup>5</sup> to detect users near the display. We also used the Microsoft Internet Explorer Web Component for the browser part of the *AmbientBrowser* system.

Figure 5 has examples of the *AmbientBrowser* system using a brightness sensor in a study room, a flow-rate sensor in a running faucet in a bathroom, and a window-embedded sensor to determine the distance from the user to the display.

Figure 14 shows the snapshot of the *AmbientBrowser* (*EnergyBrowser*) system using energy consumption by user's exercise. Both users are wearing a motion sensor that calculates the amount of energy consumed. The system renders the target Web pages incrementally in proportion to the rate at which walking or running is done. The woman at left is reading the target Web page gradually while walking/running on an Air-Walker. The woman at right is on a treadmill. Both can change the rendering speed merely by changing their walking/jogging speed.

Figure 15 shows the *AmbientBrowser* system monitoring a knife's movement in cooking in the kitchen. The system has a motion sensor attached to the knife. It renders the target Web page incrementally in proportion to the movement of the knife, such as that in cutting.

## 4 Discussion

We established some *AmbientBrowser* systems in homes and offices. From simple field tests, we found that our system provided a great deal of knowledge to people who enjoyed the Web. People could obtain unexpected knowledge from the *AmbientBrowser*.

The *AmbientBrowser* system using energy consumed during exercise to provide Web pages (see Fig. 14) is useful. Many people enjoyed exercise and browsing the Web. However, a system using energy consumed during cooking (see Fig. 15) is not useful because it is dangerous to read/browse the Web while cutting with a sharp knife. If this mechanism is to be used, the system should stop the target Web page from being rendered while cutting is being done and it should gradually render the page when the cutting task has stopped. Otherwise, we have to introduce a media conversion system from text to audio.

<sup>5</sup> <http://www.phidgets.com/>



Fig. 14. Left figure shows air-walker version. Right figure shows treadmill version.



Fig. 15. User reads/browses Web page gradually while cutting

Energy consumption, brightness, distance are useful because it is easy for users to control them. Sound volume is little useful in the alone situation. This value is not suitable in public space. However, temperature, humidity and air flow are not useful because it is difficult for users to control them. We must think about combining the situation and input. We plan to attempt various new parameters such as weight, the slider and the user's attention span.

Link navigation is a problem with the *AmbientBrowser* system. The user only controls abstract parameters. However, it should not be necessary to navigate links while jogging. The URL lists in the current implementation worked well. However, it is inefficient to manually create lists, and automatic generators of URL lists, such as read site summary (RSS) aggregators [6] [10], would be useful. We could use an RSS list as the program, showing RSS entries one-by-one in response to the users' pacing.

We found the presentation of Web pages was important after the field tests [13]. Large font sizes and appropriate typefaces were needed for legibility. Simplified

structures were also preferable for the *AmbientBrowser* system because it automatically scrolled down the display area. We found some pages where the pacing and volume of speech could be adjusted typographically (Fig. 16). These were very popular with users because incremental rendering made it appear as though the speech was real. We plan to improve the content-conversion mechanism we used to modify pages for greater legibility. We are also planning to evaluate our system.



Fig. 16. Examples of typographical tone expressions

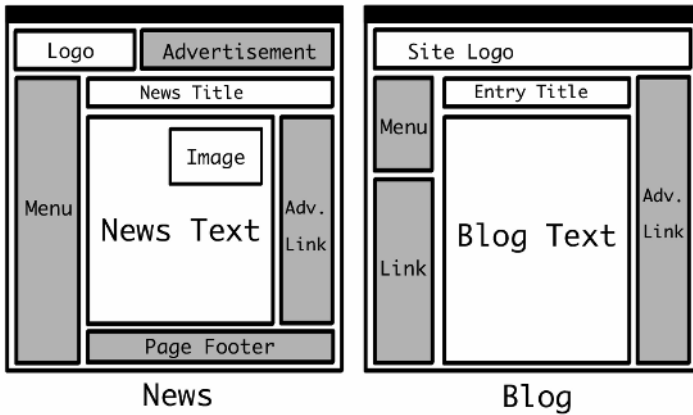
Several user comments emphasized the importance of converting Web pages for the *AmbientBrowser* system. There are many sections in these Web pages that readers do not need to see. Removing unneeded portions (such as frames, banner advertisements, and navigation links) increases readability. In addition, the font size can be enlarged to improve readability (see Fig. 17).

Variations with keywords had limitation in the prototype system because the *AmbientBrowser* only uses preset keywords. We could acquire synonyms for keyword by using a thesaurus to increase keyword variations. People's preferences may change dynamically. However, the keywords for user's preferences were static in this prototype system. We are planning to introduce a dynamic keyword-changing mechanism to locate user preferences.

*Memorium* [19]<sup>6</sup> is a persistent interface-like clock. This system provides memos and texts searched by keywords and search engines. People can use it to assist their ideas. Our system is similar to this. However, *Memorium* does not take the ubiquitous display's characteristics or text rendering method into account.

It also has problems with providing inappropriate information such as that from adult sites in public space. We have to introduce a filtering mechanism to solve this problem.

<sup>6</sup> <http://www.persistent.org/memorium.html>



**Fig. 17.** Site construction examples. Left figure illustrates construction of news site. Right figure depicts construction of weblog site.

## 5 Improvements

In response to feedback of the system, we have introduced new elements and attributes into the HTML and the content-conversion module.

### 5.1 Extension of HTML

As described earlier, gradual Web rendering controlled by the amount of calories burned requires parallel structures to be sequentially mapped within specific parameters and rendering timing to be precisely managed. Although some of these requirements can be achieved through heuristics, it is necessary for researchers to have precise control. It is also necessary to separate preprocessors and gradual renderers for implementation purposes. We therefore introduced the following elements and attributes.

**Weight Attribute.** The WEIGHT attribute defines the rendering speed, which means the amount of content rendered per unit quantity of the parameter. The WEIGHT attribute can be applied to various elements, such as DIV, SPAN, TH, TD, and IMG, but the processing methods may differ. A typical unit of the WEIGHT attribute is a byte. The WEIGHT attribute of the text defines the rendering speed. When time is assigned to the parameter, the WEIGHT attribute controls the amount of text displayed per second.

**Order Attribute.** The ORDER attribute defines the order in which the Web page is displayed. The ordering attribute can be applied to elements that define portions of Web pages. User agents that comply with gradual rendering must reorder the portions using ordering attributes according to their value when assigning a parameter to the page. The default value is “undefined”, i.e., portions that have no ordering attributes are assigned in the order of appearance within their scope, e.g., DIV and TABLE, after portions that have ordering attributes.



**Rendering Attribute.** The RENDERING attribute defines how an image is rendered, such as JPEG or PNG format. We can only use this attribute in IMG tags. We can set the following options:

- **NORMAL** (This is a conventional rendering style.)
- **ZOOM IN** (This enlarges the target image in steps.)
- **ZOOM OUT** (This reduces the target image in steps.)
- **TILE** (This divides the target image into tiles, such as  $X * Y$ . It displays the tile for the target image in steps.)
- **RESOLUTION** (This first displays the target image at low resolution, gradually displaying higher resolutions until the image becomes clear.)
- **MOSAIC** (This first displays the target image in mosaic style. It then becomes clear.)
- **LIGHTEN** (This gradually lightens the target image.)
- **DARKEN** (This gradually darkens the target image.)

In addition, the display speed (e.g., the speed of scaling up, scaling down, and lighting) is dependent on the size of the image.

**Pause Element.** The PAUSE element is used to stop gradual rendering for short periods at the position of the element. It can be regarded as an empty element, whose size is the value of the element. For example, when time is assigned to a parameter, the gradual rendering process stops for 10 seconds, and the PAUSE element has a value of 10.

**Hide Element.** The HIDE element makes its children invisible. We have defined this element to be compatible with ordinary Web browsers. The preprocessor or converter inserts HIDE elements into hide portions that are unnecessary for gradual rendering, e.g., frames and link lists.

**Clear Element.** The display is cleared when the gradual rendering process encounters a CLEAR element. This element is useful for separating a Web page into imaginary pages.

**Sound Element.** The SOUND element embeds a sound in a specific position on the page. The user agent plays the sound when it encounters a SOUND element. This is useful for attracting the user's attention. The SOUND element is a shortcut to the OBJECT element in HTML.

Figure 18 has an example of these extension elements and attributes. The header and footer are hidden using the HIDE elements. The page is rendered as follows. First, a sound is produced, and each letter in the news title is displayed, individually. Then, the news text is displayed at a speed of 20 bytes per parameter. An image for "main.jpg" is displayed more slowly than it is for "sub.jpg", because the weight of the former is half that of the latter. After the second part of the body of the news has been rendered, the display is cleared. Then, the image for "sub.jpg" and the third part of the body of the news are rendered. Figure 19 has an example of gradual image rendering.

```

<HTML>
<HEAD>
<TITLE> News Title </TITLE>
</HEAD>
<BODY>
<HIDE> Page Header </HIDE>
<SOUND src="jingle.wav">
<H1 weight="1"> News Title </H1>
<DIV id="news_body" weight="20">
The first part of the news.
<IMG src="main.jpg" weight="10">
The second part of the news.
<IMG src="sub.jpg" weight="20">
The third part of the news.
</DIV>
<HIDE> A link to the top page,
copyright, etc. </HIDE>
</BODY>
</HTML>

```

Fig. 18. Example extension elements and attributes

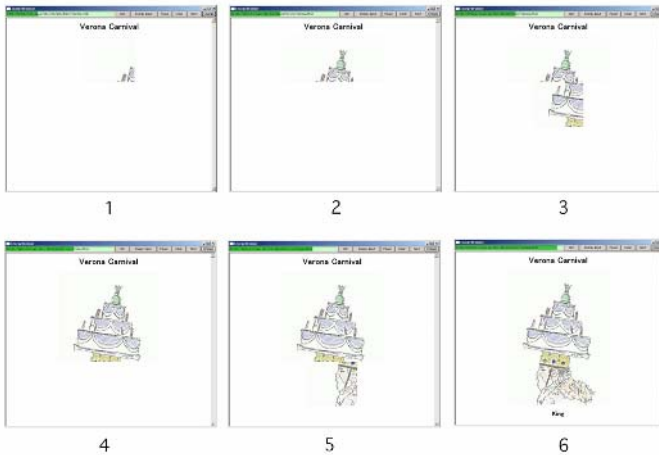


Fig. 19. Example gradual image rendering process. We used TILE (2, 3) in these figures.

### 5.3 Content Conversion Module

This module converts the fetched page into a format suitable for gradual rendering. This has two main purposes: the first is to serialize the flow of content, and the second is to stylize the page to provide better presentation during gradual rendering. This module goes through the following processes:

- Removes unnecessary parts, such as frames, banner advertisements, and navigation links.
- Enlarges the text font to improve readability.
- Fits the table to avoid horizontal scrolling when it is wider than the browser's width.
- Fits the image size to improve visibility.

- Inserts extensional elements and attributes, if heuristic rules are given. For example, if the content of the H1 elements is important, the module inserts a WEIGHT attribute with a larger value to each H1 element.
- Applies the conversion rules for content presentation to the Web page using template matching.

Figure 20 has an example of the new gradual Web rendering. Our module removed unnecessary parts and enlarged the text font and fitted the image size. It first loads the rules for essential/non-essential areas that contain the flags for essential or non-essential portions, the names of the rules, the parts of the URL that recognize the target Web content, and the start and end tag patterns (see Fig. 21).

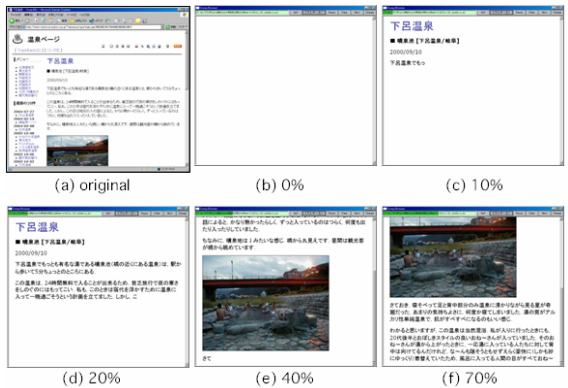


Fig. 20. Gradual-rendering process. (a) is original page.

```

Name: Remove Sample's Advertisement
Flag: Non-essential
URL: http://news.sample.site/
Start: <!-- Advertisement Start --!>
End: <!-- Advertisement End --!>

Name: Remove Sample's Menu
Flag: Non-essential
URL: http://news.sample.site/
Start: <!-- Menu Start --!>
End: <!-- Menu End --!>

Name: Main content of Sample Diary
Flag: Essential
URL: http://diary.sample.site/
Start: <!-- Diary Start --!>
End: <!-- Diary End --!>
    
```

Fig. 21. Example rule list

The system applies all rules to the target Web content. If the URL pattern for the rule corresponds to the URL for the target Web content, the system detects essential/non-essential areas by matching rules. If the matched area is non-essential, the system removes it. If the matched area is essential, the system removes all but the matched areas. Our system gives preference to non-essential rules over essential ones when targeting Web content.

The system then detects non-essential areas by using part of the URL (e.g. double-click.net). If there is a URL for an advertisement, it removes this as an advertisement area. In addition, the system counts the number of links in one area, such as in a table. If the number of links is over a preset value, it removes this area as a menu.

Non-essential parts of Web pages, such as frames, banner advertisements, and navigation links, are removed in these processes. Users can then read/browse only the target content.

## 6 Conclusion

We designed and implemented the *AmbientBrowser* system, which can be used from any location and provides Web pages to inform people on a daily basis in ubiquitous computing environment. It selects target Web pages through the utilization of various keywords and a Web search engine. It also displays these pages incrementally in relation to inputted parameters. In this paper, we explained about our system and applications. In addition, we introduced some mechanism to improve our system. People can read/browse Web pages easily yet efficiently in daily life.

We are currently evaluating the effectiveness of our system in field tests. In addition, we are also applying other types of parameters to the *AmbientBrowser* system.

## References

- [1] Dahley, A., Wisnesk, C., and Ishii, H.: Water Lamp and Pinwheels: Ambient Projection of Digital Information into Architectural Space, In Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI 1998), pp. 269- 270 (1998).
- [2] Heiner, J. M. et al.: The Information Percolator: Ambient Information Display in a Decorative Object, In Proceedings of CHI '91, pp. 85-90.
- [3] IBM Corp., "ViaVoice Dictation, ViaVoice Developers Toolkit." <http://www-4.ibm.com/software/speech/>.
- [4] Igarashi T. and Hughes J. F.: "Voice as Sound: Using Non-verbal Voice Input for Interactive Control," In Proceedings of 14th Annual Symposium on User Interface Software and Technology, ACM UIST'01, Orlando, FL, pp. 155-156, (2001).
- [5] Ishii, H. et al.: Pinwheels: Visualizing Information Flow in an Architectural Space, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2001), pp. 111-112 (2001).
- [6] Luke H., SharpReader, <http://www.sharpreader.net/>
- [7] Maslow, A. H.: Motivation and Personality. Harper & Row, New York (1970).
- [8] McCarthy, J., Costa, T. and Liongosari, E.: UniCast, OutCast & GroupCast: Three Steps Toward Ubiquitous, Peripheral Displays, In Proceedings of the third International Conference of Ubiquitous Computing (UbiComp 2001), pp. 332-345 (2001).

- [9] NEC Corporation, "SmartVoice", <http://www.amuseplus.com/product/voice/>.
- [10] Nakamura S., WeBoX, <http://www-nishio.ist.osaka-u.ac.jp/~nakamura/webbox/>.
- [11] Nakamura S., Minakuchi M., and Tanaka K.: AmbientBrowser: Web Browser in Life, Ambient Intelligence and (Everyday) Life, pp. 83- 92 (July 2005).
- [12] Nakamura S. Minakuchi M., and Tanaka K., EnergyBrowser: Walking in the World Wide Web, 11th International Conference on Human-Computer Interaction (HCI 2005), Vol. 2 No. 42 (July 2005).
- [13] Nakamura S., Minakuchi M. and Tanaka K.: Energy Browser: To Make Exercise Enjoyable and Interesting, ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2005), pp.258-261. (June 2005).
- [14] Starner T., Auxier J., Ashbrook D. and Gandy M: The Gesture Pendant: A Self-illuminating, Wearable, InfraredComputer Vision system for Home Automation Control andMedical Monitoring. Proc. ISWC 2000, International Symposium. on Wearable Computers (ISWC 2000), pp. 87-94 (2000).
- [15] Stasko, J., Miller T. Pousman, Z., Plaue C., and Ullah O.: Personalized Peripheral Information Awareness through Information Art, In Proceedings of the 6<sup>th</sup> International Conference on Ubiquitous Computing (UbiComp 2004), pp. 18-35 (2004).
- [16] Tanaka, K., Nadamoto, A., Kusahara, M., Hattori, T., Kondo, H., and Sumiya, K.: Back to the TV: Information Visualization Interfaces Based on TV-Program Metaphors. In Proceedings of the IEEE ICME2000, pp. 1229-1232, 2000.
- [17] Tsukada, K. and Yasumura, M.: Ubi-Finger: Gesture Input Device for Mobile Use, Proceedings of APCHI 2002, Vol. 1, pp. 388-400 (2002).
- [18] Tsukada, K. and Yasumura, M.: Ubi-Finger: Gesture Input Device for Mobile Use, Companion Proceedings of UbiComp'2001, Technical Report: GIT-GVU-TR-01-7 (2001).
- [19] Watanabe K. and Yasumura M.: Memorium: The Concept of a Persistent Interface and its Prototype, WISS 2002 (in Japanese).
- [20] Weiser, M.: The Computer for the Twenty-first Century, Scientific American, pp. 94-104 (1991).

# Online Music Search by Tapping

Geoffrey Peters, Diana Cukierman, Caroline Anthony, and Michael Schwartz

Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada  
{gpeters, diana, canthony, mpschwar}@sfu.ca  
<http://www.songtapper.com>

**Abstract.** Query by Tapping is an emerging paradigm for content-based music retrieval, which we have explored through our web-based music search system. Based on the results obtained from our system we argue that searching for music by tapping the rhythm of a song’s melody is intuitive and effective. In this paper we describe two novel algorithms to analyze tapping input. We present results indicating good accuracy rates among a broad spectrum of both trained and untrained users. Query by tapping has an important potential as a form of human-computer communication. We indicate how our algorithms to analyze tapping might be used in other areas such as music education and user authentication.

## 1 Introduction

One theme of Ambient Intelligence is concerned with creating ubiquitous and intuitive user interfaces to intelligent systems [1]. The act of tapping for the purpose of communication is an intuitive one, dating back to the age of Morse code and the telegraph. However, in the modern computing paradigm, rhythmic tapping is not currently a mainstream method of human-computer interaction. The current most popular methods of user input, namely the acts of “typing” and “clicking” by using the keyboard and mouse, are adequately suited for text and image based interaction, but are not very effective for content-based music information retrieval. We believe that tapping has a great potential for a rediscovery as a useful form of communication, especially as a way to express queries in music search. This so-called Query by Tapping (QBT) is an emerging paradigm for content-based music retrieval that has only just begun to be explored [23] [6] [20].

We developed an online QBT system that allows web site visitors to search for music by tapping the rhythm of song melodies. We developed two novel<sup>1</sup> rhythmic encoding algorithms which capture the essential elements of rhythm, allowing it to be represented in textual form. Thanks to these encoding algorithms (which we describe in Section 5), the system is tolerant of tempo variations and errors in the input. The system can also be trained by users to recognize new songs.

---

<sup>1</sup> One of these algorithms, the Rhythmic Contour String algorithm, was first presented by our team at AAI-05, as described in [20].

Another important advantage of our system is that it does not require any special hardware or software. The input device is the space bar on the computer keyboard, and the software is compatible with most web browser client platforms. The naturalness and accessibility of the web based interface has allowed us to gather and analyze feedback from a broad spectrum of Internet users, with varying levels of musical ability.

In Sections 2 and 3, we present an overview of previous work in tapping-based music retrieval, and a general background relating QBT to other forms of Music Information Retrieval (MIR). In Section 4 we describe the design and implementation of our web-based QBT system, including an automatic learning module allowing users to train the system. In Section 5 we then discuss the mechanics of two novel algorithms for analyzing tapping. In Section 6 we analyze user feedback and usage data collected from our web-based QBT system to provide some promising results concerning the effectiveness of the system. Based on our experimentation and analysis of the data obtained, we conclude that tapping as a form of human-computer communication has an important potential for various future applications.

## 2 Tapping as a Natural Act

An underlying theme behind Ambient Intelligence is that computers and humans should interact in a way that is unobtrusive and natural [1]. This involves an intuitive user interface as well as enabling the computer system to have enough information available to make appropriate decisions in order to communicate effectively with the user.

Across all cultures, people move their bodies to the rhythms of music, by clapping, tapping, drumming, singing, dancing, or rocking an infant [22]. As we have discovered through a study of user experiences with our QBT system, the act of tapping to express a musical rhythm is surprisingly easy to many people, even if they are not musically trained. While pitch-based approaches such as humming or whistling have potentially more descriptive capability for song melodies, they are not always as intuitive as tapping, nor are they as accurate for musically untrained or tone deaf users.

The concept of clapping along to a song is very natural. Many children learn to play clapping games such as patty-cake. At music concerts, audiences sometimes spontaneously clap in a specific rhythm, along with the musicians on stage. In addition to clapping their hands, people commonly use other physical motions to follow rhythms, such as foot stomping, knee smacking and finger snapping.

The task we propose, through our QBT system, is for users to tap the rhythm of a song's melody on their computer keyboard's space bar, in order to retrieve a list of songs which have melodies that contain a similar rhythm. Because of the underlying rhythmic encoding algorithms our system employs, users tap at their own tempo and pace. Overall tempo variations such as speeding up or slowing down do not produce adverse effects. Users may begin and end tapping anywhere within the song, and errors in the input are tolerated.

Other QBT systems, which were developed independently, use different input methods (as well as different underlying algorithms). Jyh-Shing et al's system [23] allows the user to tap a rhythm on the end of a Karaoke microphone, and uses a matching algorithm based on timing vectors and dynamic programming. Eisenberg et al's system [6] requires the user to tap at a certain tempo on a specialized drum pad (using either the hands or drumsticks), and uses an algorithm based on MPEG-7 beat vectors. Jyh-Shing and Eisenberg both provide evidence to support the claim that beat information is an effective feature for song search in a music database. The QBT system that we developed extends the tapping human-computer interface concept further, in terms of ease-of-use and accessibility, by being the first QBT system to be deployed as a web application. Our encoding and matching algorithms produce comparable, if not superior performance to existing systems, and also have the advantage of relative simplicity and ease-of-implementation.

The rhythm of a song's melody is often equivalent to the rhythm of the words in the lyrics, as the song is commonly sung. Each syllable or two of a word might represent one beat. Whether or not the user remembers the specific words, he or she can often recall and reproduce (to some extent) the rhythm and most likely the tone of the melody. We have found from user observation that tapping the melody of a song is quite easy to do, if the user sings aloud and taps at the same time.

### 3 Overview of Music Information Retrieval

Music Information Retrieval (MIR) is a field that is concerned with the general problem of searching for music in a library based on the content of the music, rather than on textual meta-data such as the song title or artist name. Users may express queries based on features obtained from existing recordings (such as in Query by Example), or through inputting a performance of the rhythm or pitch of a song's melody (such as in Query by Tapping or Humming respectively).

The motivation behind research in MIR extends to both academic and commercial applications. Melucci et al [17] emphasize that music is an important form of cultural expression, and with increasing digital access, librarians need more effective methods for organizing and retrieving it. Kosugi et al [13] describe a Karaoke machine that lets users select the song they want by singing a part of it.

Query by Example, as previously mentioned, describes the situation where a user would provide the system with a music recording, and the system would extract features from the recording to allow the discovery of other music with similar features. Examples of such as systems are discussed in [10] and [11].

In contrast with Query by Example, another area of MIR is based on user-input, where users search for a song by some element of its musical content, using a natural input method such as tapping, humming, or singing. The overriding focus in user-input based MIR is on the melody of songs, since that is the musical component most easily identifiable by musically untrained users [17]. In addition to QBT, which utilizes the rhythmic features of a melody, other input methods



are Query by Humming or Singing, which utilize the pitch features of a melody [8] [14] [12]. There are also systems which incorporate both rhythm and pitch [5].

Some user-input based MIR systems rely on extended knowledge from the user, such as musical training or the ability to read music [6]. Ideally, the user interface to a MIR system should be so intuitive that the user does not need special training. As we further analyze in Section 6, some users of our QBT system did not have any musical experience, and were still able to achieve reasonable success rates.

Another common feature of MIR systems is a tolerance to various types of user input errors. This is especially important if users have varying levels of musical ability. The way errors are handled depends on the algorithm which is used to analyze the songs and find matches. Techniques such as dynamic programming, n-grams, and approximate string matching have been used in previous work [26]. Kline & Glinert [12], as well as Kosugi et al [13] focused on trying to improve robustness against specific types of user errors.

In our QBT system, by focusing on the rhythmic features and ignoring pitch features, we need not worry about errors in pitch. Some users could be tone-deaf, or just incapable of playing the song with the correct pitch. For dealing with errors in rhythm, our system employs a fast approximate matching algorithm which we describe in Section 5.1.

Scalability of MIR systems is another important feature. As Bainbridge et al [2] describe, accuracy in matching is not the only scalability concern; we also should worry about the run-time complexity of the algorithm in general. They claim existing algorithms based on approximate matching are too slow and impractical for databases larger than 10,000 songs, when a linear scan of the database is used. They describe a Bioinformatics heuristic approach called BLAST, which provides a database indexing scheme that they purport can do much better. In order for this, and other general scalability questions, to be answered, further empirical testing is needed.

A key challenge with many MIR systems is the task of creating a song database that has an appropriate music representation format. Many MIR systems represent melody and rhythm in secondary formats, which cannot be easily determined from digital audio recordings that are commonly found in music libraries. User-input based approaches rely on higher-level music representation forms such as MIDI (Musical Instrument Digital Interface) which specify precise note and timing information for the melody, instead of working directly on a digital audio signal. Although attempts at pitch and rhythm extraction from audio recordings have been made [9], the current state of the art cannot accurately extract melodies and rhythms from the majority of audio recordings. Unfortunately, this means that databases containing higher-level music representations need to be created manually, or through techniques such as Optical Music Recognition [2].

Our web-based QBT system attempts to overcome the lack of available song data by allowing web visitors to train the system by tapping new songs, and then associating the tapping data with specific song names. This user-driven training approach presents some challenges of its own, but is part of a growing number

of web-based Artificial Intelligence systems that take advantage of the “collective mind” of Internet users [3] [4] [28]. We describe this user-driven training mechanism in Section 4.3.

Web-based QBT systems (such as ours) can be deployed using the hardware and software that already exists on the majority of Internet-enabled workstations. Without any special hardware, our QBT system allows users to ‘tap’ using the space bar on their computer keyboard. Cross-platform compatibility is achieved by software deployment using web standards such as HTML, Java Applets, and Macromedia Flash, which can be accessed from most modern browsers such as Firefox or Internet Explorer. Our web-based QBT system provides a MIR system that is nearly universally accessible, through its intuitive interface and cross-platform compatibility.

## 4 Our Application: Web-Based Query by Tapping

We developed a web based system for music search by tapping, using the algorithms described in Section 5. The initial prototype system allowed searching of a database of 30 children’s songs, and collected feedback from the users about the success of their tapping experience and search results. The subsequent second-generation system allows users to expand the song database by training the system themselves. In both systems, the user taps the rhythm of a song’s melody on his/her computer keyboard and the tapping sequence is recorded using a Java applet (or Macromedia Flash applet, as in the second-generation system) and then is sent to our application server for analysis. The search results are displayed in the browser, and the user has the opportunity to give feedback on the outcome of the search.

By using existing, commonplace web-based standards and technologies for application delivery, as well as the ubiquitous standard computer keyboard for user input, we have allowed a broad audience of Internet users to access our system, who may not be experts in either music or computer technology. Access to this larger audience, provided with our online feedback system, has allowed us to gain new insights into the applicability and usefulness of a QBT system on a larger scale than a controlled laboratory environment.

### 4.1 User Interface

A user can visit our web site [21], and use a Java or Flash applet to tap a rhythm of a song, using the space bar on the keyboard (see Figure 1). The applet will generate a MIDI file and automatically upload it to our application server, which will display the search results in the browser. This allows users to input rhythmic data without having special equipment such as a MIDI keyboard or a microphone device. The system relies on users having sufficient musical skill to input a rhythm that can be accurately discriminated. As we describe in Section 6 we have found that many untrained users do possess enough rhythmic skill to use the system. By using a web based interface, users can be located in diverse parts of the globe, and still interact with our software.

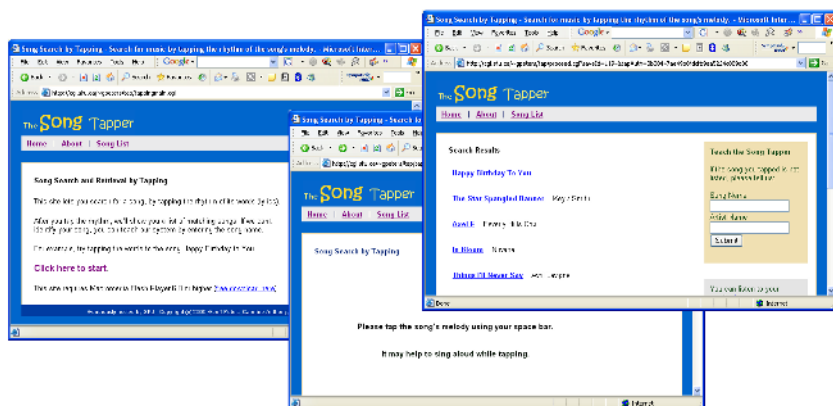


Fig. 1. Screenshot of Online Song Search by Tapping

## 4.2 Architecture

Our software makes use of the MIDI (Musical Instruments Digital Interface) standard to represent recordings of user search tapping sessions. A musical performance is encoded as a series of events, with each event having properties such as the time, duration, and velocity of the note. When a user begins tapping, a new event would be recorded for each time the space bar is pressed on the keyboard. A special MIDI keyboard is not required for our tapping application, as we have utilized the standard alphanumeric computer keyboard as a tapping device. Our client-side browser applet, originally implemented in Java and now also implemented in Flash, allows a MIDI file to be generated from a tapping performance on a computer keyboard. On the server side, our Perl-based software makes use of the Perl::Midi library to parse and analyze MIDI files, which are passed through our encoding algorithms (as described in Section 5) to generate strings that represent the rhythm. The Perl String::Approx library is used to carry out fast approximate string matching. In addition, a MySQL database is used to store the song data.

## 4.3 Web Based Training Module

One problem, as previously identified, is the lack of a large database of songs which can be used as training data for our system. We are currently overcoming this difficulty by allowing the visitors of the site to train our system in real time. Our database of songs grows incrementally thanks to training performed by volunteer visitors.

Here is a listing of steps in the process that this training occurs:

1. *User Input of Song Rhythms*

To search, the user taps the rhythm of a song using our web site tapping applet (by tapping the space bar on the computer keyboard). A MIDI file is generated automatically, and the song is uploaded to the application server.

## 2. *Data Processing, Search, and User Feedback*

If the system guesses the song correctly then the user gives positive feedback indicating that this is the case, and the user's data is added to candidate rhythms for that song.

When subsequent searches are done the software searches through all candidate rhythms for all songs for a more accurate search.

If the system does not guess correctly then the name of the song is entered by the user (if known), and new training data is added to the database. To facilitate more accurate training, the system shows a list of songs with similar names that are already in the database (by doing a metadata search), allowing the user to add the data to an existing song or create a new song.

Some users naturally have more skill at training the system than others, due to their musical ability. In the near future, a feature will be added whereby regular users will be able to sign in, and if they generally input more accurate rhythms then their input will be given more weight when the search is done.

The data in the database which matched incorrectly (when searching) is tracked and if it accumulates too many incorrect matches then it is flagged as possible bad data. The system also tracks good matches.

This training module, which we have recently implemented and launched on our web site, has already begun to increase the number of songs in our song database. As the database grows, it will become more useful for users, hopefully attracting more users to train the system. We expect that we will reach a point where we will be able to correctly identify a good portion of the rhythms that are tapped into the system. Once the database has reached such a useful state, it should be possible to connect our system to an online music store, or even make it accessible on mobile devices to take advantage of the growing mobile/wireless music market.

## 5 Our Algorithms for Query by Tapping

We present two algorithms, the Rhythmic Contour String algorithm and the Phrase String algorithm. These algorithms encode rhythm in textual string form, and specify how these strings should be compared using approximate string matching to determine rhythmic similarity.

Our algorithms are designed to analyze a monophonic tapping sequence (that is, a sequence in which only one key is ever depressed at any one time). When a tapping session is performed by a user, we record the time at which the user depresses and releases the key for each tap. These key presses are represented in a sequence of note onset times  $o_i$  and release times  $r_i$ . In our analysis, we are only concerned with the timing of when the user strikes the keyboard for each tap. We are not at all concerned with the timing of when the user releases each key, assuming the user must release the key in order to strike the next tap. The exception is the last note in the sequence, for which we do consider the release time because there is no subsequent note to follow.

In this discussion, we consider a beat's duration to consist of the time between the onset of the current tap to the onset of the next tap. An entire tapping session, then, consists of a sequence of  $n$  beat durations  $d_i$  where  $d_i = o_{i+1} - o_i$ , except for the last  $d_n$  which takes the value of  $d_n = r_n - o_n$ .

In order to allow for global tempo independence, we normalize the durations of the beats. To normalize the durations, the average duration of a beat is calculated, and then each beat's duration is divided by the average duration. For a particular song that is being analyzed, the graph of the normalized duration can be plotted per beat, as beats progress through time. By comparing these "duration plots" for various songs, it can be observed that most songs have unique "duration functions". Figure 2 shows the duration plot for a user's performance of the first part of the song 'Are You Sleeping'. The corresponding sheet music notation is shown in Figure 3.

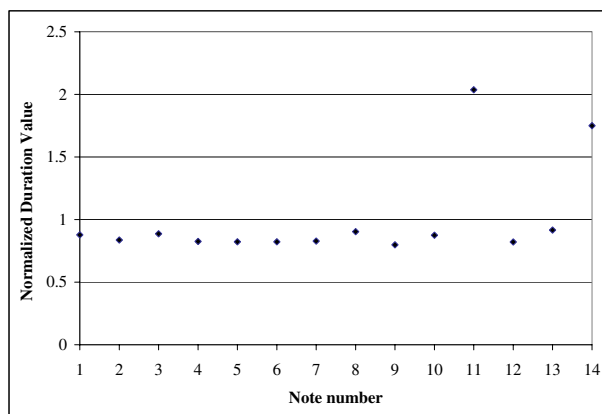


Fig. 2. Normalized Duration Plot of 'Are You Sleeping'

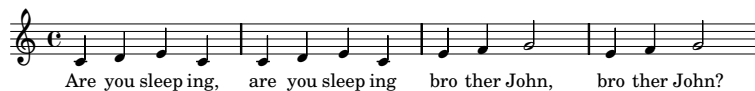


Fig. 3. Sheet Music Notation for 'Are You Sleeping'

## 5.1 Rhythmic Contour String Algorithm

Our goal is to encode the tapping duration information  $d_i$  as a string of characters, so that approximate string matching can be used, leveraging fast existing algorithms (such as Sun and Manber's algorithm [25]). To achieve the string encoding, we introduce the idea of a Rhythmic Contour. We conceive the idea of a rhythmic contour as a description of how the duration of each consecutive beat is different from the previous one. The rhythmic contour is calculated by

---

**Algorithm 1.** Encode Rhythmic Contour String  $str$  from  $n$  Beat Durations  $d_i$

---

**Require:**  $d_i$  is an array of normalized beat duration values where  $i$  takes values from 1 to  $n$ .  $T$  is a similarity threshold value  $> 0$  such as 0.25.

**Ensure:**  $str$  is a string of length  $n - 1$  consisting of any of the characters s, u, and d.

```

1:  $str \leftarrow$  empty string
2: for  $x = 1$  to  $n - 1$  do
3:    $c \leftarrow d_{x+1} - d_x$ 
4:   if  $|c| < T$  then
5:     append character 's' to string  $str$ 
6:   else if  $c < 0$  then
7:     append character 'd' to string  $str$ 
8:   else
9:     append character 'u' to string  $str$ 
10:  end if
11: end for
12: return  $str$ 

```

---

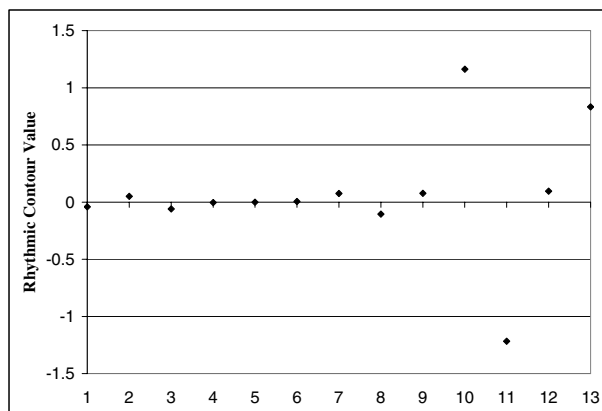
subtracting the duration of each pair of consecutive notes. Thus we calculate the contour values  $c_i$  as  $c_i = d_{i+1} - d_i$ .

The result of our Rhythmic Contour String encoding for a song of  $n + 1$  beats is a string with  $n$  characters, with the three symbols 's', 'd', and 'u' (meaning the duration stays the same, goes down, or up, respectively). Our encoding algorithm produces a character for each sequential contour value  $c_i$ . For each contour value  $c_i$ , if  $|c_i| < T$  (where  $T$  is the value of a threshold such as 0.25), we append the character 's' to the string. Failing that, if  $c_i < 0$  we append the character 'd' to the string, or if  $c_i > 0$  we append the character 'u' to the string.

For example, a song that has  $n$  beats of equal value would have  $n-1$  calculated contour values, each with a value of 0, and therefore a string of  $n-1$  's' characters. A simple song that has only two beats, the last with a much longer duration than the first, would have a single, positive contour value, and a string with simply the character 'u'. It may be illustrative to examine a plot of the Rhythmic Contour values (see Figure 4) and associated Rhythmic Contour String ('ssssssssudsu') for the example song, and compare it to the plot of duration for the same song (see Figure 2).

A related approach to analyzing rhythm was presented by Doraisamy and Ruger [5], where ratios between note onset times were examined. They observed that by comparing the differences between consecutive note onset times, exact beat and measure information does not need to be known, and quantization on a predetermined base duration is not required. They also remark that the usage of note onset times was earlier explored by Shmulevich et. al [24].

**Approximate Matching Procedure.** Each song to be put in the database is pre-processed by the above algorithm and a Rhythmic Contour String is generated for each. When a query is received from the user, the user's input rhythm is analyzed, and a Rhythmic Contour String is generated for it as well. Now that both the user's input rhythm and the rhythms of the songs in the database



**Fig. 4.** Rhythmic Contour Plot of ‘Are You Sleeping’. These rhythmic contour values would produce the string ‘ssssssssudsu’.

are represented as strings, we use an approximate string matching procedure to determine the closest matching songs. A fast approximate string matching algorithm as specified by Sun and Manber [25] is used to calculate the edit distance between the input string and each string in the database<sup>2</sup>.

The edit distance (also known as the Levenshtein measure [15]) is defined as the minimum number of transformation operations needed to transform one string into another string. Transformations can be composed of one or more of the following three operations:

1. removing a character
2. inserting a character
3. substituting one character for another

The edit distances between the input string and each string in the database are sorted into ascending order. The song’s string in the database with the least edit distance to the input rhythm string is considered to be most probable match.

**Similarity to Parsons Code.** Our Rhythmic Contour String encoding is similar to a code developed by Denys Parsons in 1975, called the Parsons Code [19]. Our approach uses a string to describe the contour of how a melody’s *rhythm* changes through time, whereas the Parsons Code describes the contour of how a melody’s *pitch* changes through time. Parsons created a dictionary of musical themes, where he used the melodic contour of a song (the way in which a melody goes up or down on the music scale/staff) to encode a string with the letters ‘U’, ‘D’ or ‘R’ for up, down, and repeat, respectively. For example, the Parsons Code for the beginning of “Twinkle Twinkle Little Star” is RURURDDRDR-DRD. Our Rhythmic Contour String encoding is essentially a Parsons Code for rhythm.

<sup>2</sup> A freely available Perl implementation of this algorithm can be found in the Perl String::Approx library, at <http://search.cpan.org/>

Some researchers have used the Parsons Code as a basis for a MIR system, such as Tseng's implementation [26] which uses melodic (pitch-based) contours. Another MIR application which uses Parsons Code can be accessed at Musicpedia [18]. The Musicpedia application allows the user to input the Parsons Code directly, or to hum into a microphone to generate the Parsons Code. For some users, direct input of a Parsons Code or Rhythmic Contour String may seem unnatural or unintuitive. Thus, our system does not require the user to input a Rhythmic Contour String directly, but rather generates the appropriate string based on the user's tapping performance.

**Scalability of the Algorithm.** The string encoding procedure in Algorithm 1 does not pose a scalability concern since the number of beats is relatively small, and the number of operations is linear to the number of beats. Songs that are added to the database only need to be encoded once, and the encoded strings are stored for future use.

For the approximate string matching algorithm, as described by Sun and Manber [25], the matching of the encoded query string with a single song in the database takes  $O(nk\lceil m/w \rceil)$  where  $n$  is the size of the string in the database,  $k$  is the number of errors allowed in the input,  $m$  is the length of the encoded query string, and  $w$  is the system word size (for example: 32 bits). Interestingly, because of the dependence on system word size in the matching algorithm, on a system with a 32 bit word size, queries that have between 34 to 65 taps take twice as long as queries that have up to 33 taps. Queries with over 65 taps take at least three times as long. That is because a session with  $m + 1$  taps generates a contour string with  $m$  characters; thus, for example, a query with 34 taps generates a string with 33 characters which exceeds the system word size of 32.

This approximate string matching for a query in our system is currently performed for each song in the database, in a linear scan. Thus the search time grows linearly with the number of songs in the database. As Bainbridge et al [2] describe, such a linear scan poses a scalability concern when the database is on the order of 10,000 songs. Some possible ways to allow the database to scale further would be to use a parallel server configuration to divide up the load of the query on to multiple servers, or to develop a partial indexing scheme (borrowing ideas from the BLAST Bioinformatics system mentioned by Bainbridge et al), which would reduce the number of songs that would need to be examined for a particular query. As previously mentioned, future work needs to be done to address these scalability concerns.

## 5.2 Phrase String Algorithm

We have also developed a Phrase String algorithm to be used in conjunction with the Rhythmic Contour String algorithm. We first sort the song matches by the edit distance obtained by the Rhythmic Contour String algorithm, and then further sort by the edit distance obtained by the Phrase String algorithm. Using these two algorithms in conjunction appears to produce even more accurate results than using a single algorithm alone. Further empirical testing is currently in progress.



---

**Algorithm 2.** Encode Phrase String  $str$  from  $n$  Beat Durations  $d_i$ 

---

**Require:**  $d_i$  is an array of normalized beat duration values where  $i$  takes values from 1 to  $n$ .  $n \geq 1$ .

**Ensure:**  $str$  is a string consisting of the characters ‘1’ and ‘p’.

```

1:  $str \leftarrow$  empty string
2: for  $x = 1$  to  $n$  do
3:   append character ‘1’ to string  $str$ 
4:   if  $\text{IsEndOfPhrase}(d_x)$  then
5:     append character ‘p’ to string  $str$ 
6:   end if
7: end for
8: return  $str$ 

```

---

The Phrase String algorithm is similar to the Rhythmic Contour String algorithm in that both algorithms encode a string based on a sequence of duration values  $d_i$  and then use approximate string matching to determine the edit distance between a query string and strings in the database.

The Phrase String algorithm generates a “phrase string” which contains the symbol ‘1’ for each beat in the song. Additionally, for each beat which denotes an end of a musical phrase, the symbol ‘p’ is inserted after the ‘1’ for that beat. The phrase detection algorithm which we have implemented identifies beats with durations that are longer than the two closest neighboring beats, as possible ends of phrases. This phrase detection approach is a much simplified version of the algorithm described by Melucci and Orio [17]. For example, the Phrase String for the duration values shown in Figure 2 would be ‘1111111111p11p’. In this example, there are 14 characters of ‘1’, since there are 14 beats. A ‘p’ is inserted after the eleventh ‘1’ because that beat is longer in duration than the two neighboring beats. A ‘p’ is always added at the end of the string, because we expect that the user would likely stop tapping at the end of a phrase.

Approximate string matching can again be used on this Phrase String, with a slight modification to the approximate string matching algorithm. Instead of allowing three different transformation operations to calculate the edit distance, only the following two are allowed:

1. removing a character
2. inserting a character

The transformation operation that is not directly allowed is the substitution of a character for another. The reason for this restriction is that the symbol ‘1’ denotes a beat, and ‘p’ denotes an end-of-phrase, and these concepts are not obviously substitutable. In the Rhythmic Contour String algorithm we do allow substitutions because each character represents a transition from one beat to another, and a substitution simply implies that one of the notes had an incorrect duration. But for the Phrase String, a substitution implies that the user made an error where he/she intended to play a longer note, but instead

played an extra note, or vice versa. We consider this to be an error worth an edit distance penalty of 2 instead of 1 (that is, it is accounted for by an insertion and a deletion). Any mismatches between the strings will be detected by the two allowed transformation operations. Thus this algorithm still allows for errors such as the insertion of an extra beat, or a missing end-of-phrase symbol.

## 6 Results and Analysis

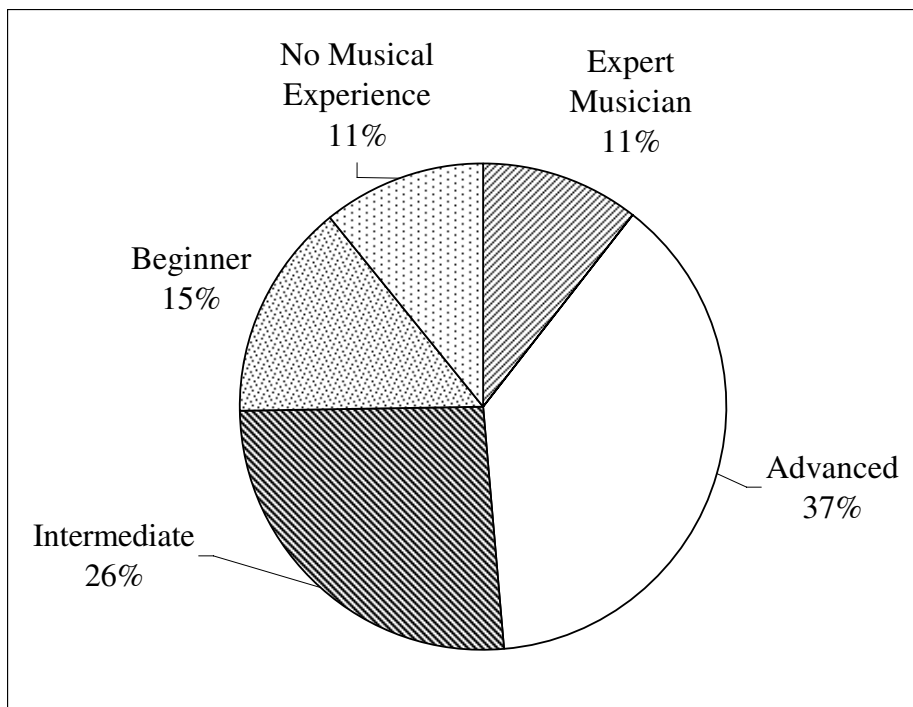
Our initial prototype system, for which the results will be analyzed in this section, made use of the Rhythmic Contour String algorithm as described above, and had a database of 30 children's songs. It did not make use of the Phrase String algorithm, which was developed after the initial prototype was created. However, the Phrase String algorithm is implemented in our second generation system, which is currently collecting data from users who access it.

For the initial prototype system, users were asked to provide feedback on whether the song they tapped was identified as the first song in the search results, or failing that, if the song was in the top ten results. Users were also asked to self-rate their musical ability on a scale of 1 to 5. The users' tapping sessions were recorded and data was collected such as the duration of each session, and the number of notes tapped.

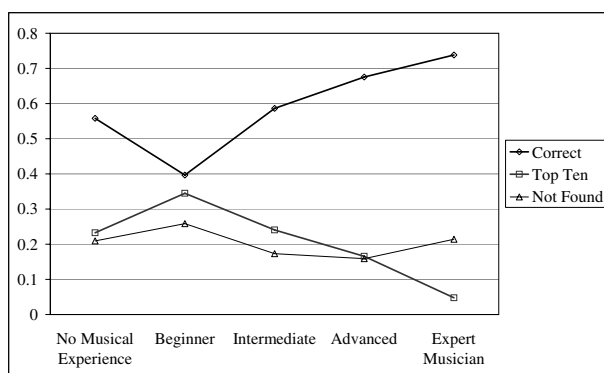
Users gave feedback for 518 tapping sessions, performed from May to September 2005. Regardless of user ability level, 62% of all feedback reports indicate a first place ranking, that the system determined the correct song as the first ranked search result. 20% of all feedback reports indicated that the correct song was in the top ten, but was not the first ranked result. 18% of all feedback reports indicate the correct song was not present in the top 10 search results.

Figure 5 shows the breakdown of the self-rated ability levels of the users in the collected data. Figure 6 shows that the proportion of first place ranking outcomes is largest for experts, with a proportion of 73.8%. Beginners have the lowest first place ranking outcome proportion, of 39.7%. This data suggests that musical ability of users plays a significant role in the overall usability and effectiveness of the system, although beginners can still experience success. Surprisingly, those who rated themselves as having 'no musical experience' did better than the beginners, with a first place ranking proportion of 55.8%.

Figure 7 relates the accuracy rate to the number of notes tapped in the session. The accuracy rate is calculated as the proportion of the sessions which had a first place ranking outcome. One can observe that sessions with fewer than 10 notes tapped had an accuracy rate of zero, and from this point onwards, the accuracy rate appears to increase linearly with number of notes tapped, until 25 notes are tapped. After 25 notes tapped and onward, the accuracy rate appears to remain fairly constant. The oscillation observed from 50 notes and above is likely due to a lower concentration of data with these number of notes. The range of this graph encompasses 98.26% of the feedback data, as the number of sessions with 60 notes tapped and greater were not numerous enough to provide an interesting graph beyond that point.

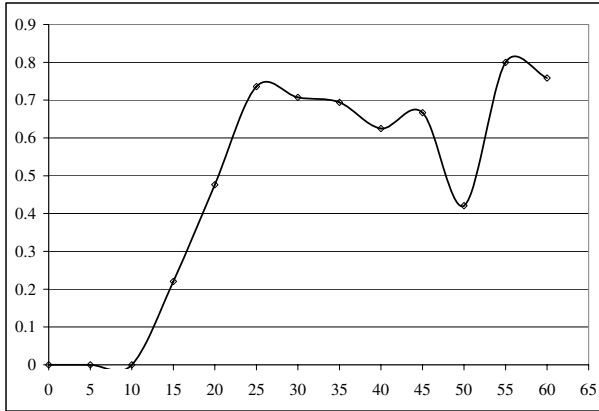


**Fig. 5.** Self-Reported Ability Levels of Users. For each of the 518 tapping sessions, users were asked to self-rate their musical ability level.



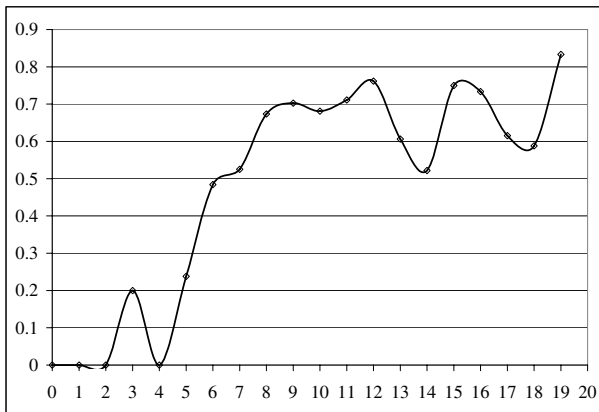
**Fig. 6.** Breakdown of Outcome by Ability. The vertical axis shows the proportion of total searches that had the specified outcome, for the particular user ability level.

Figure 8 relates the accuracy rate to the length of the session in seconds. As before, the accuracy rate is calculated as the proportion of the sessions which had a first place ranking outcome. A positive correlation is apparent



**Fig. 7.** Accuracy Rate Versus Number of Notes Tapped

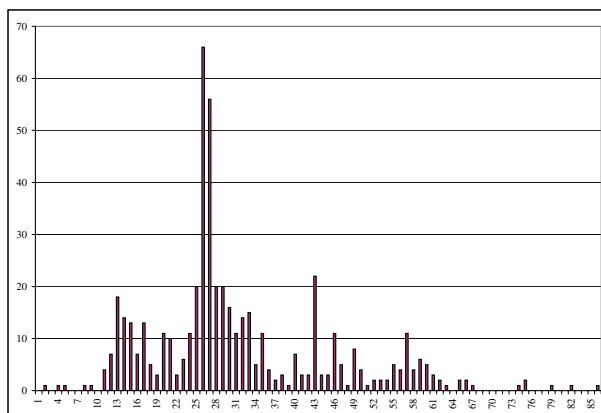
until the session length reaches 8 seconds. From this point onwards, the accuracy rate appears to remain fairly constant. The range of this graph encompasses 90.7% of the feedback data, as the number of sessions with length 20 seconds or greater were not numerous enough to provide an interesting graph beyond that point.



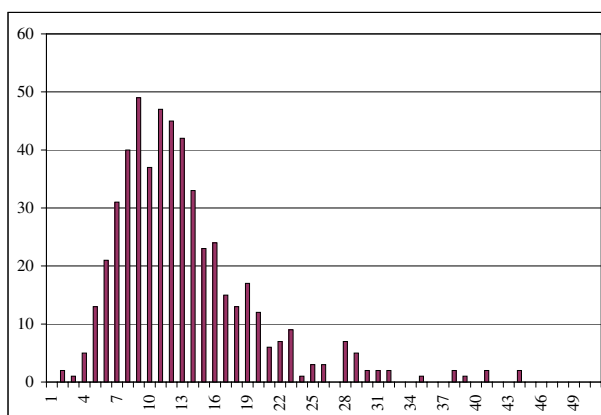
**Fig. 8.** Accuracy Rate Versus Session Length in Seconds

When accessing our system, users were not given any instructions as to how many taps they should make, or how long their tapping session should take. Users were free to tap at their own tempo, and were not restricted with a metronome. The distribution of sessions by number of taps is shown in Figure 9 and has a mean of 30.8 taps and a standard deviation of 15.2. The distribution of sessions by length of session in seconds is shown in Figure 10

and has a mean of 11.8 seconds and a standard deviation of 6.75. We observed that on average, users tend to tap enough notes to provide the system with adequate information for a successful search. Notice that users enter an average of 30.8 taps (Figure 9) and that the system performs the best when at least 25 taps are entered (Figure 7). A similar observation can be made connecting results in Figure 10 and Figure 8.



**Fig. 9.** Histogram of Sessions by Number of Taps. The mean number of taps is 30.8 and the standard deviation 15.2.



**Fig. 10.** Histogram of Sessions by Length of Session. The mean session length in seconds is 11.8 and the standard deviation 6.75.

It is expected that if the database grows to several thousand songs, the Rhythmic Contour String algorithm alone will no longer provide enough information to discriminate song content accurately. However, the combination of the Rhythmic Contour String algorithm with the other algorithms (such as

the Phrase String algorithm described above), will likely increase the discriminatory ability of the system.

## 7 Other Applications of Query By Tapping

Computer-aided music education is one possible application area of our tapping algorithms, which we believe merits future exploration. We conducted an informal experiment at AAAI-05 [20] where about 50 subjects attempted to tap a song using our system, and those who had difficulty in tapping the correct rhythm were asked to observe an expert musician tapping the same song. By learning through imitation, the subjects were able to improve the accuracy of their tapping on a subsequent attempt, as evidenced by a smaller edit distance to the intended song. In this experiment, rather than using the tapping interface as a search tool, it was instead used as a way to assess the subject's musical knowledge of a particular song. Such a tool could be useful in music classrooms to provide immediate feedback to students who are learning new rhythms.

Some other application areas, which are inspired by the themes of Ambient Intelligence, are to integrate tapping interfaces into mobile and embedded applications. For example, if a tapping sensor were integrated into the home, a user could simply tap a certain rhythm on a table-top to cause the lights to dim, and a romantic song to begin playing on the home stereo system. If a tapping interface were integrated into a children's toy, a child could communicate with the toy by tapping a certain song, such as 'Old Macdonald Had a Farm', causing the toy to respond by playing back a lively recording of the same tune. If integrated on a mobile device, a user could tap the rhythm of a song on his/her cell phone and the phone's software could immediately download a ringtone that contains a similar rhythm.

Another intriguing possible application would be to use tapping as a way to identify a person, in a situation where security is not a major concern. We propose to call this application area 'Authentication by Tapping'. For example, imagine that everyone in a household has a personalized rhythm, which could be fairly short. A child could configure the door to his/her room to only unlock to his/her 'secret knock' (but the parents would naturally have an override feature).

## 8 Conclusion

Through illustrating our online music search system that utilizes Query by Tapping, we have argued that tapping is a highly promising paradigm for human-computer interaction. By creating a web-based Query by Tapping system, and making it freely accessible on the Internet, we have been able to expose our system to a broad spectrum of users. We have presented good evidence from our study to show that tapping is an intuitive and effective means of communication, that can be used by people with varying musical abilities. We have presented

two novel encoding algorithms for rhythm, the Rhythmic Contour String algorithm and the Phrase String algorithm, which are effective enough to merit use in future applications which utilize tapping as a means of user input, such as in music education, mobile devices, or home integration.

While our approach to rhythmic searching allows humans to use a very natural and intuitive input method, it enables the computer system to make decisions that humans might find difficult. Untrained humans can mimic or shadow the clapping of another human, and can clap out the rhythm of a song they already know, but sometimes have a hard time when asked to identify a song based on a monotone rhythm they hear. As future work, we propose an experiment whereby the computer system's skill at recognizing songs based on rhythm would be compared to humans. While such an experiment would be akin to a sort of limited Turing Test for rhythm (similar to a Feigenbaum test [7]), which in value is debatable [16], it might also be interesting to compare our system's performance against both expert musicians as well as untrained humans. In any case, the broad range of potential applications involving tapping as a means of human-computer interaction shall make Query by Tapping an attractive focus of future study.

*Acknowledgements.* Special thanks to the SFU School of Computing Science, the SFU Faculty of Applied Science, the SFU Faculty of Business Administration, the SFU Faculty of Arts, and the Simon Fraser Student Society for providing funding and support of this project, to allow us to present our demonstration at AAAI-05 in Pittsburgh in July 2005. Also thanks to series editor Yang Cai from Carnegie Mellon University for his interest in our work.

## References

- [1] Aarts, E.: Ambient Intelligence: A Multimedia Perspective. *Multimedia*, IEEE. **11**, **1** (2004) 12–19
- [2] Bainbridge, D., Nevill-Manning, C., Witten, I., Smith, L., and McNab, R. Towards a Digital Library of Popular Music. *Proceedings of the Fourth ACM Conference on Digital Libraries*. ACM Press: NY. (1999)
- [3] Chklovski, T., and Gil, Y.: Towards Managing Knowledge Collection from Volunteer Contributors. *Proceedings of AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC05)*. (2005)
- [4] Chklovski, T.: 1001 Paraphrases: Incenting Responsible Contributions in Collecting Paraphrases from Volunteers. *Proceedings of AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC05)*. (2005)
- [5] Doraisamy S., and Ruger, S.: An Approach Towards a Polyphonic Music Retrieval System. *Proceedings of the Second Annual International Symposium on Music Information Retrieval: ISMIR 2001*. (2001)
- [6] Eisenberg, G., Batke, J-M., and Sikora, T.: BeatBank - An MPEG-7 Compliant Query by Tapping System. 116th AES Convention, Berlin. (2004)
- [7] Feigenbaum, E.: Some Challenges and Grand Challenges for Computational Intelligence. *Journal of the ACM (JACM)*. **50**, **1** (2003) 32–40

- [8] Ghias, A., Logan, J., and Chamberlin, D.: Query by Humming - Musical Information Retrieval in an Audio Database. *ACM Multimedia 95 - Electronic Proceedings*. (1995)
- [9] Gomez, E., Klapuri, A., Meudic, B.: Melody Description and Extraction in the Context of Music Content Processing. *Journal of New Music Research*. Routledge. **32**, 1 (2003) 23–40
- [10] Haitsma, J., and Kalker, T.: A Highly Robust Audio Fingerprinting System. *International Symposium on Musical Information Retrieval (ISMIR2002)*. (2002) 144–148
- [11] Harb, H., and Chen, L.: A Query by Example Music Retrieval Algorithm. *4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS03)*. (2003) 122–128
- [12] Kline, R., and Glinert, E.: Approximate Matching Algorithms for Music Information Retrieval Using Vocal Input. *Proceedings of the Eleventh ACM International Conference on Multimedia*. ACM Press: NY. (2003)
- [13] Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., and Kushima, K.: A Practical Query-by-Humming System for a Large Music Database. *Proceedings of the Eighth ACM International Conference on Multimedia*. (2000) 333–342
- [14] Kosugi, N., Nagata, H., and Nakanishi, T.: Query-by-Humming on Internet. *Lecture Notes in Computer Science*. Springer-Verlag. **2736** (2003) 589–600
- [15] Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* (Feb. 1966) 707–710.
- [16] Luger, G.: *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison-Wesley. **5th ed.** (2005) Ch 1
- [17] Melucci, M., and Orio, N.: Musical Information Retrieval Using Melodic Surface. *Proceedings of the Fourth ACM conference on Digital Libraries*. (1999) 152–160
- [18] The Open Music Encyclopedia. Online Resource. <http://www.musicpedia.com> (2005)
- [19] Parsons, D.: *The Directory of Tunes and Musical Themes*. Spencer Brown. (1975)
- [20] Peters, G., Anthony, C., and Schwartz, M.: Song Search and Retrieval by Tapping. *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*. (2005) 1696–1697
- [21] Peters, G., Anthony, C., and Schwartz, M.: The Song Tapper. Online Resource. <http://www.songtapper.com> (2005)
- [22] Phillips-Silver, J., and Trainor, L.: Feeling the Beat: Movement Influences Infant Rhythm Perception. *Science*. **308** (2005) 1430–1430
- [23] Jyh-Shing, R., Hong-Ru, L., and Chia-Hui, Y.: Query by Tapping: A New Paradigm for Content-Based Music Retrieval from Acoustic Input. *Lecture Notes In Computer Science*. **2195** (2001) 590–597
- [24] Shmulevich, I., Yli-Harja, O., Coyle, E., Povel, D.-J., and Lemstrom, K.: Perceptual Issues in Music Pattern Recognition Complexity of Rhythm and Key Finding. *Proceedings of the AISB 99 Symposium on Musical Creativity*. (1999) 64–69
- [25] Sun, W., and Manber, U.: Fast Text Searching: Allowing Errors. *Communication of the ACM*. **35(10)** (1992) 83–91
- [26] Tseng, Y.: Content-based Retrieval for Music Collections. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press: NY. (1999)
- [27] Uitdenbogerd, A., and Zobel, J.: Manipulation of Music For Melody Matching. *ACM Multimedia 98*. (1998) 235–240



- [28] Von Ahn, L., and Dabbish, L.: Labeling Images with a Computer Game. ACM CHI. (2004)
- [29] Wang, Y., Kan, M., Nwe, T., Shenoy, A., and Yin, J.: LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics. Proceedings of the 12th Annual ACM International Conference on Multimedia. ACM Press: NY. (2004)

# Whistling to Machines

Urko Esnaola and Tim Smithers

Project MiReLa,  
San Sebastián Technology Park,  
Mikeletegui Pasealekua, 53  
20009 Donostia - San Sebastián, Spain  
{uesnaola, tsmithers}@fatronik.com

**Abstract.** The classical approach to improve human-machine interaction is to make machines seem more like us. One very common way of doing this is to try to make them able to use Human Natural Languages. The trouble is that current speech understanding techniques do not work well in uncontrolled and noisy environments, such as the ones we live and work in. Nor do these attempts mean that the machines use our languages in the way we do: they typically don't speak much like we do, and we mostly have to speak to them in special unnatural ways for them to be able to understand. Rather than require people to adapt how they speak to machines, so that the machines can understand them, we present a simple artificial language, based upon musical notes, that can be learned and whistled easily by most people, and so used for simple communication with robots and other kinds of machines that we use in our everyday environments.

## 1 Introduction

As the devices and machines that we add to our living places, work places, and public places, have become “intelligent,” or “smart,” or “clever,” so the complexity and associated difficulty of using and interacting with them has increased. In the home, this trend started with the programmable video recorder (VCR), [14], but has been replicated in many other devices and machines since. The widely adopted, or classical approach to solving these human-machine interaction difficulties, is to try to make the machines seem more like us; more human-like. A common means to do this, is to give the machines so called Natural Language interfaces, so that we can talk to them, and they can talk back, in (one of) our own languages. In this paper we present a different approach to human-machine interaction. This uses a simple artificial language, based upon musical notes, that can be easily learned and whistled, or produced using other simple musical instruments or modern mobile telephones, and which works in the kinds of uncontrolled and noisy environments in which we live and work.

The design and use of this artificial language is motivated by both practical and philosophical concerns. On the practical side, we need a simple and reliable means of communication with machines in uncontrolled and often noisy conditions. Natural Language technologies for speech recognition are not yet good

enough for these kinds of challenges—the best systems work well only under well controlled situations: a directional microphone placed close to the speaker’s mouth, clearly spoken phrases or single words, with little or no significant background noise of any kind. On the philosophical side, we do not believe that machines should be made to appear to use any of our sophisticated Natural Languages. It only appears that these machines speak and understand our language. They typically do not really use our language, they do not or cannot listen to and understand what they say, and for these systems to work, it is often we (humans) who have to adapt how we speak and what we say, so that the machines understand us—which we do not count as proper Human Natural Language use.

In the next section we review some related work on using musical sounds in human-machine interfaces and communications. In section 3, we present three illustrative examples of using musical sound as a simple human-machine communication language. Section 4 presents the structure and pragmatics of our musical language, and section 5 explains the simple agent-based sound recognition and language processing system we use. Section 6 then presents some of the real-world trials that have been used to evaluate the research and development of the language, and in a final section, we present a brief discussion and some conclusions.

## 2 The Sound of Music

Sound, has always been an important medium for communication between people, including non-speech sounds like music. In Europe, horns have long been used for the transmission of different kinds of messages over long distances, and bells have long been used, and still are, for communicating the passing of time. In other cultures, drumming is a popular means of communication. But sound, especially musical sound, seems to be rather neglected as a medium for human-machine communication. In the case of human-computer interaction, it’s more a case of having been forgotten. As [27] comment, many early computer users took advantage of an auditory output to know better what the computer was doing. One quite common trick was to tune an AM radio to the RF “noise” produced by the circuits of the computer CPU, and thereby listen to what it was doing. The sound patterns produced—a kind of machine music—were quite easily learned and recognised, and helped programmers to diagnose and even debug their programs.

Smoliar, [26], argues that communication is an act of intelligent behaviour, and that by looking at music, rather than natural language, we can more clearly focus on this idea of communication as a behavioural process, and thus how it should fit with and work with other behaviours. Alty, [3], further highlights the advantages of using music as a medium of communication:

“Music is all-pervasive in life and forms a large part of people’s daily lives. It is very memorable and durable. Most people are reasonably familiar with the language of music in their own culture. Once learned, tunes are difficult to forget.”

Taking this as a starting point, Vickers and Alty, [27], present a modern version of the early “machine music,” in a program debugging support system that uses music, in its normal full sense, to provide well structured, rich, but easily learned and recognised information about a program’s execution flow. They argue that this kind of auralisation of information offers benefits over and above the visualisation offered by conventional graphical user interface techniques.

Earcons (ear-cons, as opposed to eye-cons = i-cons = icons), see [5], are abstract synthetic sounds that can be used in structured combinations to create sound messages that represent different parts of an application interface. Brewster, [8, 9], has tried to promote the use of musical earcons in computer interfaces. He and his colleagues has sonified several interfaces, and have shown them to be effective at communicating complex information in sound. In choosing to investigate music-based earcons, Brewster et al, [19], point out that, rather than depending upon the sound structure of (non-musical) earcons to transmit information, music can transmit information without requiring that its structure is understood explicitly. This pioneering work has, however, seen little further application, despite proving to be an effective way of improving and extending human-computer interfaces, and not just for visually impaired users.

Earcons and debugging music are examples of machines making sounds to users, but what about users making sounds to the machines? Which is what is needed for effective inter-communication. As we have mentioned, the Classical approach to this side of the human-machine interaction equation, is to try to use a Human Natural language. The long history of speech recognition research shows this to be complex, difficult, and still not solved well enough for anything but simple dialogue systems working in well controlled conditions. An interesting alternative approach, using whistling, has been promoted and demonstrated by Mark Bohlen, an artist, with his Universal Whistling Machine (UWM), [6]. Here is how Bohlen describes his work:

“UWM is an investigation into the vexing problem of human-machine interface design. Whistling is much closer to the phoneme-less signal primitives compatible with digital machinery than the messy domain of spoken language. As opposed to pushing machines into engaging humans in spoken language, UWM suggests we meet on a middle ground. Whistling occurs across all languages and cultures. All people have the capacity to whistle, though many do not whistle well. Lacking phonemes, whistling is a pre-language language, a candidate for a limited Esperanto of human-machine communication.”

—*Marc Bohlen, The Universal Whistling Machine, 2004*

The music language we present here shares much with Bohlen’s idea that, in looking for good forms of human-machine interactions, we should look to meet the machines half way, “on a middle ground,” rather than try to drag them towards us—which, all too often, actually results in us having to go a long way over to them: typically having to unnaturally distort our Natural Language use in the process. Though whistling was not the original form of production of the music

language we present, it has become the production means of choice, and the one for which the language is now designed, because whistling is something most people can do relatively easily, and is common to all known human communication cultures. It is even the basis of several human languages, such as “Silbo,” that still survives (just) in Gomera; the smallest of the Canary Islands, off the West coasts of Africa, [15], where it is used for long distance communication, across nearby hill tops.

Furthermore, by using musical notes to define the alphabet of our language, natural intervals to form words, and a range of notes that can easily be whistled by most people, we end up with phrases being short musical tunes that are, easily produced, easily learned and remembered, and so easily recognised, and understandable without explicitly having to interpret the note structure of the words and phrases—which would otherwise make the language production and understanding harder.

### 3 Music for Machines

In this section, we briefly describe three examples of the use of our music-based human-machine communication language. The first, a mobile robot called MiReLa, has been the principle motivation and focus for our music language research and development. The second example describes our attempts to enable the robot MiReLa to operate the lifts (elevators) in the building in which it works; an example, this time, of machine-machine communication, but of a type people can also understand. The third example illustrates the use of the music language to control a desk lamp; a simple device of the type we find in everyday living and working environments.

#### 3.1 The Robot MiReLa

MiReLa is a mobile robot (based upon a RWI (Real World Interface) B21 system) that is the centre piece of Project MiReLa, a research project that has been conducted at the San Sebastián Technology Park, since September 1997, see Figure 1.

The two basic aims of Project MiReLa are, robust navigation in semi-structured uncontrolled environments, and effective people-robot interaction in service robot scenarios. Two important characteristics of this project are that it has adopted a purely Behaviour-based approach to developing a robust and reliable navigation competence for the robot, [24, 12], and that it has been conducted, from the start, in a real place, *not* in a laboratory: a real place where other people work, public events take place, often large changes occur as part of its normal usage, and where there is “musak” (background piped music) everywhere.

MiReLa has been developed to act as a kind of artificial guide, that can take people from the main entrance of the building to other parts of the building, such as the bar (on the ground floor), auditorium or seminar rooms (on the first



**Fig. 1.** Robot MiReLa

floor), or to people's offices (on the second and third floor), [11]. To aid and support this role, MiReLa uses an artificial language based upon words made up of musical notes: and this is where its name *Mi-Re-La* comes from (the notes e, d, A). Originally, we designed this language, which we call MiReLa Music Language (MML), for our own use, as researchers working with the robot, to have a convenient way to communicate with the robot: both to tell it to do things, and have it tell us things. Later, we saw that other people, not associated with the project, were able to learn to use this language, and able to whistle it quite easily.

Using MML, it is possible to tell MiReLa to go to particular places in the building, and it is possible to tell it to stop and start while it is navigating. MiReLa uses the same language to communicate: to say what it is about to do, or to tell people other things about what it is doing, and about what state it is in. It can ask for a door it needs to pass through to be opened, for example, [13]. We will see more details and examples of MiReLa Music Language in section 4, but the use of this MiReLa Music Language in the development and many demonstrations of the robot, have shown its use to be both reliable and robust in real conditions, as well as quite easy for other people to learn and use successfully. Some of these real-world trials are described in section 6.

### 3.2 The Lifts

MiReLa does not have arms to manipulate things with. It cannot physically open doors, or operate the lifts (elevators), for example. As a first step towards making

it possible for MiReLa to use its Music Language to do this, we have installed a system that uses Dual Tone Multi Frequency (DTMF) sounds—like telephones use—to control one of the lifts in the main building of the San Sebastián Technology Park. This installation uses DTMF boards (NORCOMM NC400 [22]) which offer several relay controls, together with pairs of microphones and speakers installed in the control panels of the lift on each floor, and one inside the lift cabin. This set up allows MiReLa to call the lift, by generating the sound of the appropriate DTMF code, to hold the lift doors open, so that it can get into and out of the lift, and to tell the lift which floor number it needs, once it is inside the lift cabin. The lift is also able to generate DTMF sounds to indicate (to MiReLa or to people) that it has detected and understood a request, and when the lift has arrived. This simple system has proved to be relatively easy to design, install, and make work reliably, despite the large variation in acoustics between the spaces in front of the lifts at each floor, and in the lift cabin. We now plan to extend this installation to use the full MiReLa Music Language described in section 4.

### 3.3 The Desk Lamp MiFaRe

A system, using the same language and signal processing techniques developed for MiReLa, has also been implemented to control a desk lamp, used in one of our offices. This system consists of a board containing eight relays ([20]) which can be controlled from a PC over a standard serial port, together with a microphone and loudspeaker connected to the sound card in the PC. When the musical name for the lamp, MiFaRe, is detected and recognised, a signal is sent to the relay to change state: to switch the lamp on, if it was off, or to switch it off, if it was on. In each case, the system responds by also saying “okay,” not just by turning on or off the light. Once again, this system has proved to be both reliable and robust in a normal office environment, with typical kinds and amounts of environmental noise from computer ventilation systems, people, music, telephone conversations, etc.

## 4 MiReLa Musical Language

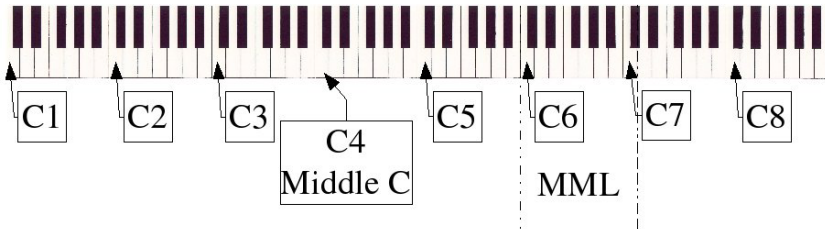
Experiments have shown that a healthy young person can hear all sound frequencies from approximately 20Hz to 20,000Hz [10]. A typical person, however, can sing in only a rather narrow range of frequencies. Table 1 shows the ranges demanded by the different voices of classical opera, for example, [16].

*MiReLa Music Language* (MML) has been designed to use notes from the C major scale: from 1,046.5Hz to 2,093Hz. This is C6, note *c*, octave 6, following the octave designation system of the American Acoustic Society, [4], to C7, note *c*, octave 7. Figure 2 indicates where on a standard piano keyboard this MML range is. This frequency range is higher than most humans can reach by singing so as to avoid interference from people (particularly children) speaking (and singing), yet, it is the frequency range most people can whistle in. Based

**Table 1.** Human singing voice ranges, in Hz

| Voice     | Range in Hz       |
|-----------|-------------------|
| Bass      | 87.31 – 349.23    |
| Baritone  | 98.00 – 392.00    |
| Tenor     | 130.00 – 493.88   |
| Contralto | 130.81 – 698.46   |
| Soprano   | 246.94 – 1,174.70 |

upon some trials conducted with different people, we concluded that people can whistle up to a frequency of about 2,093 Hz (C7) without great effort, with higher being increasingly harder to whistle. And they can whistle down to a frequency of about 783 Hz without great effort (G5), but lower notes become harder to whistle with high power. The notes in the frequency range C6-C7 can also be played on many simple musical instruments, such as Xylophones, Harmonicas, Treble Recorders, Txistu (a traditional Basque flute played one-handed), and can also be produced using many modern mobile telephones.



**Fig. 2.** A standard piano keyboard indicating the note names following the octave designation system of the American Acoustic Society, [4], with the MML alphabet notes indicated at octave C6-C7

Most musical sounds are formed from a combination of a fundamental tone plus a number of harmonics. The harmonics have frequencies an integer number of times the frequency of the fundamental tone, which often (but not always) has the largest amplitude. The timber of the instrument depends on the amplitude of the harmonics [23]. In the case of whistling, the fundamental frequency is easily given the highest amplitude in production. (Though, with practice, it is possible to change this.) This property of whistled sounds is useful for reliable recognition of the musical notes in MML.

#### 4.1 The MML Alphabet and Phrases

MML is a proper language, albeit an artificial one, because it has a well defined grammar, lexicon, dictionary, as well as some simple rules for constructing words, and for combining them into produceable phrases. It is also a proper language



**Table 2.** The MML alphabet note names and frequencies. The first column shows the names of the notes in the standard British notation, the second column shows the notes in the standard Tonic Sol Far names, [17], the third column shows the notes using the American Acoustic Society system, [4], and the fourth column presents the note frequencies in Hertz.

| Notes |     |    | Frequency |
|-------|-----|----|-----------|
| c     | Do  | C6 | 1,046.52  |
| d     | Re  | D6 | 1,174.64  |
| e     | Mi  | E6 | 1,318.52  |
| f     | Fa  | F6 | 1,396.92  |
| g     | Sol | G6 | 1,567.98  |
| A     | La  | A6 | 1,760.00  |
| B     | Si  | B6 | 1,975.52  |
| C     | Do  | C7 | 2,093.04  |

because both the people *and* the machines that use it, must both hear and recognise it, as well as produce it correctly. The alphabet used to form the words of MML is formed by the eight notes shown in table 2.

Each statement, or sentence, in MML starts with the name of the machine to which it is directed, unless this is clear from the context. The structure of a MML sentence can take one of the following forms:

[Name] + Verb + Arguments  
 [Name] + Adverb  
 [Name] + Interjection

where:

**Name** = the name of the machine to which the sentence is directed;  
**Verb** = the action to be performed;

**Arguments** = needed parameters values for the action, which can be Nouns, or Adjectives;

**Noun** = predefined names of things, such as robot transitions, and numbers;

**Adjective** = indicates a quality of a noun;

**Adverb** = “yes”, or, “no”, or “OK”; and

**Interjection** = “hello”, or “bye”.

Communication from a machine to one or more persons follows the same structure, except that machines does not add the name of the person at the start of the phrase (which typically they do not know, nor have a way of knowing). Instead a machine may add its own name at the beginning or not, depending on the context.

**Word Separation.** The two tonic notes *c* (C6) and *C* (C7) are special. They serve to separate the words in MML phrases. Thus words in MML may be

composed of any notes in the alphabet except  $c$  and  $C$  which, if heard or detected, indicate the word has finished. Also, a silence after a detected note, indicates the end of a word as well. Thus note  $c$  in MML is the sound of silence.

**Phrase Termination.** A phrase is terminated with the tonic note  $c$  or with a silence. When the note  $c$  or a silence is detected, depending on the words detected before, it is possible to determine if it is the end of the phrase or not.

If a long silence is detected, even if the phrase is not a complete phrase, it is considered to be terminated. If it is not a complete, legal MML phrase, the statement is ignored. For example, if *a name + a verb* are recognised and then no more note detections occur, though they are expected the phrase is considered to be terminated, and it is ignored as it is incomplete.

**Note Duration.** The duration of the notes doesn't affect the meaning of the words they form. That's to say, there is no restriction on the rhythm or meter of the MML tunes. This makes it easier to play them on any instrument or to whistle them and get them to be recognised.

Even if the duration of the notes doesn't affect the detection of the *MiReLa Music Language*, a uniform way was chosen to generate the language by machines. It was decided to use all notes of the same duration to generate words in MML, and make the silences to last double the duration of a note. Thus, in the examples presented in the next sections, semi-crotchets are used to represent the notes which form the words in the *MiReLa Music Language*, and crotchets for the separation notes  $c, C$ .

## 4.2 From Musical Sounds to MML Phrases

The vocabulary for MML is defined and stored in a text file, as a set of triples: word; category; musical note sequence. It is used to translate sequences of notes into words. Figure 3 shows an example of a vocabulary item.

To improve the reliability of note detection and recognition, all the consecutive repetitions of a note are considered as one note. So, the sequence  $c c c$  is detected and recognised, as one note  $c$ . This helps to make the recognition more reliable in places with significant acoustic echo, as is the case where the lifts are located in the Main Building of the San Sebastián Technology Park. This, of course, means that MML words cannot be defined using repeated notes.

In MML, Names are composed of any three non-repeating notes, except the top or bottom tonic  $c$ . Verbs, Adverbs, Nouns and Interjections may be composed of

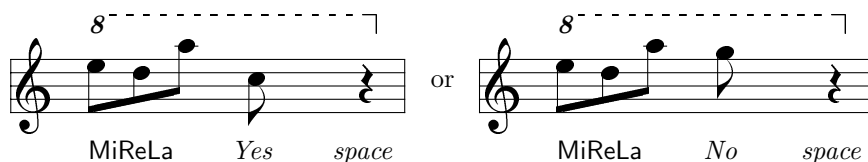
```
{
MiReLa // Name of a machine
NOUN   // Category
edA    // Note sequence
}
```

**Fig. 3.** An example of a vocabulary item in MML

any non-repeating number of notes, except top or bottom tonic  $c$ . The adverbs “Yes” and “No” are commonly used in reply to questions. To simplify the use of these words, two special case rules have been defined, which are applied to the sound heard in response to a question needing an answer of type “Yes” or “No”:

Name +  $c$  means “Yes” , and  
Name +  $g$  means “No”.

For example to answer “Yes” or “No” to a question from the robot MiReLa you can whistle or play:



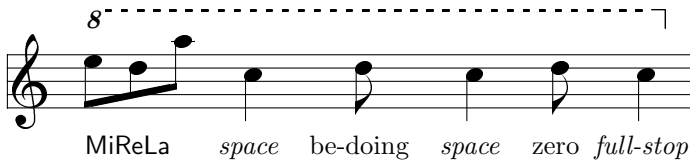
### 4.3 Directed Listening

Knowing the structure of the phrases and the musical composition of the words, it is possible to focus the sound signal processing on trying to detect certain specific notes, rather than trying to look for all possible notes in the sound signal. For example, in the case of the robot MiReLa, all the phrases directed to it start with its name which is composed of the notes  $Mi$ ,  $Re$ ,  $La$  or  $e$ ,  $d$ ,  $A$ , so the robot initially only tries to detect a note  $e$ . Any other notes played are ignored until an  $e$  is detected. After it detects an  $e$ , it tries to detect the  $d$  that follows, and then the  $A$  that should come next in its name. Having recognised its name, the robot waits for a  $c$  or a silence. It then looks in the vocabulary for the defined Verbs, Adverbs and Interjections that can follow a name, and focuses the sound signal processing on only the possible notes that start a Verb, Adverb, or Interjection. After a verb has been recognised, it focuses its attention on the notes that could form possible arguments to the verb, if any.

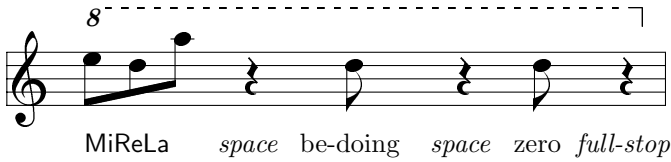
This technique filters the note detection at a low level. It is similar to what we humans do when listening to a spoken language. In human communication, there is a very close relation between what we hear and what we think we are listening to, and so to what we expect to hear, based upon our inherent knowledge of the language or languages we use [1].

### 4.4 MML in Action

To tell the robot MiReLa to go to a particular place, it is told to “be doing” a particular action set, referred to as a *Transition Group*. This makes MiReLa execute the series of *Transition Groups* it needs to complete, before it can be doing the nominated *Transition Group*, see [12] for more details. For example, to tell MiReLa to be doing Transition Group zero in MML, we play or whistle:



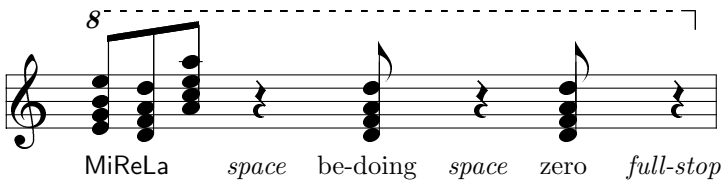
or, instead of using a tonic note, *c* or *C* (the octave above), to mark the word spaces, we could use a rest (silence) as follows:



or any combination of rests and tonics, whatever makes for the easier production of the MML phrase. Similarly, the *full-stop* at the end of the phrase can be marked by a tonic *c* or *C*.

It can be seen from this example that the Transition Group zero is only denoted by the number zero, the note *d*, since the verb be-doing can only take, as an argument, a *Transition Group* number. Furthermore, we can see from this example that the note *d* is used both to denote the verb be-doing and the number zero. Once again, the correct interpretation of these two one note words depends upon its place in the full phrase: only a verb, an adverb or an interjection can follow a name, so, in this case the first *d* is a verb (be-doing) and not a number, and the second *d* must be a noun argument for the preceding verb, not a verb, an adverb or an interjection. Verbs, adverbs, and interjections, on the other hand, must all have different note sequences, since any of them can come after a Name. Similarly, nouns and adjectives must also have distinguishable note sequences, since they can be arguments of verbs.

To the request to be doing *Transition Group* zero, MiReLa replies saying:



from which it is seen that MiReLa has a different voice; one that speaks in chords (a homophonic voice), rather than in single notes (a monophonic voice), unless it is speaking to another machine. Also, MiReLa uses silences for spaces and full-stops, not the tonic notes *c* or *C*.

### 4.5 An Example MML Conversation

Here we illustrated the use of MML in a conversation between MiReLa the robot and a lift in a building: a machine to machine interaction. It shows how MML is used to bind both the agents involved—MiReLa and the lift—into an effective interaction, and to mediate the transitions from one situation to the next in the overall scenario. It also illustrates how the contexts of the situations and the MML exchanges makes it possible for any people involved in the scenario to follow what is going on between the two machines, without necessarily completely understand what each machine is saying.

In this example, MiReLa speaks in a monophonic voice (single notes, not chords) since it is speaking to another machine, as does the lift, *LiftOne*. The musical notation here is also un-metered (it has no time signature). This is because the rhythm and meter are free in the production of MML tunes, though both MiReLa and *LiftOne* are programmed to produce MML phrases with a constant meter.

The example scenario is composed of a sequence of six situations, as follows.

**Situation 1:** MiReLa arrives at the area in front of the lifts, on it's own, or with one or two people ...

*MiReLa says:* LiftOne call\_lift [Name + Verb]  
*LiftOne says:* Yes [Interjection]

**Situation 2:** MiReLa (with person or people) waits for the lift to arrive ...

*LiftOne says:* liftOne has-arrived [Noun + Verb]  
*MiReLa says:* LiftOne hold-doors-open [Name + Verb]  
*LiftOne says:* Yes [Interjection]

Musical notation for Situation 2. The top staff (LiftOne) contains notes for "LiftOne", "hold-doors-open", and "full-stop". The bottom staff (Yes) contains notes for "Yes" and "Yes". There are "8" markers above the first and last notes of each staff.

**Situation 3:** MiReLa enters the lift (with person or people) ...

*MiReLa says:* LiftOne got\_to\_floor 2 [Name + Verb + Argument]

*LiftOne says:* OK [Interjection]

Musical notation for Situation 3. The top staff (MiReLa) contains notes for "LiftOne", "go-to-floor", "space", "two", and "full-stop". The bottom staff (LiftOne) contains notes for "Yes" and "Yes". There are "8" markers above the first and last notes of each staff.

**Situation 4:** Lift, with MiReLa (with person or people) arrives at the second floor ...

*LiftOne says:* liftOne has-arrived [Noun + Verb]

*MiReLa says:* LiftOne hold\_doors\_open [Name + Verb]

*LiftOne says:* Yes [Interjection]

Musical notation for Situation 4. The top staff (MiReLa) contains a single note. The bottom staff (LiftOne) contains notes for "LiftOne", "space", "has-arrived", and "full-stop". There is an "8" marker above the first note of the bottom staff.

**Situation 5:** MiReLa exits the lift (with the person or people) ...

*MiReLa says:* LiftOne bye [Name + Interjection]

*LiftOne says:* bye [Interjection]

**Situation 6:** MiReLa departs from the lift area, on its way to where it needs to go (with the person or people) ...

## 5 A Society of Agents

The implementation of the signal processing and semantic processing of the MiReLa Music Language is based upon a collection of simple agents that form a network. The design and organisation of these agents have been inspired by Marvin Minsky’s concept of *The Society Of Mind*, [21], which, in Minsky’s words,

“... tries to explain how minds work. How can intelligence emerge from non-intelligence? To answer that, we’ll show that you can build a mind from many little parts, each mindless by itself.”

— *The Society Of Mind*, 1 PROLOGUE pp. 17

Minsky continues by defining these “little parts” as follows:

“ I’ll call “Society of Mind” this scheme in which each mind is made of many smaller processes. These we’ll call agents. Each mental agent by itself can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence.”

— *The Society Of Mind*, 1 PROLOGUE pp. 17

The word agent here is thus used with the same meaning that Minsky gives to it. And the agents that we present in the implementation below, are designed to be simple entities: they are not able to engage in negotiations, nor are they able to reason about their own actions, or the actions of other agents. As Minsky says

“... we should not try to find a close analogy between the low-level agents of a single mind and the members of a human community. Those tiny mental agents simply cannot know enough to be able to negotiate with one another or to find effective ways to adjust to each other’s interference.”

— *The Society Of Mind*, 3.2 NONCOMPROMISE pp. 33

## 5.1 Putting Agents Together into a Society

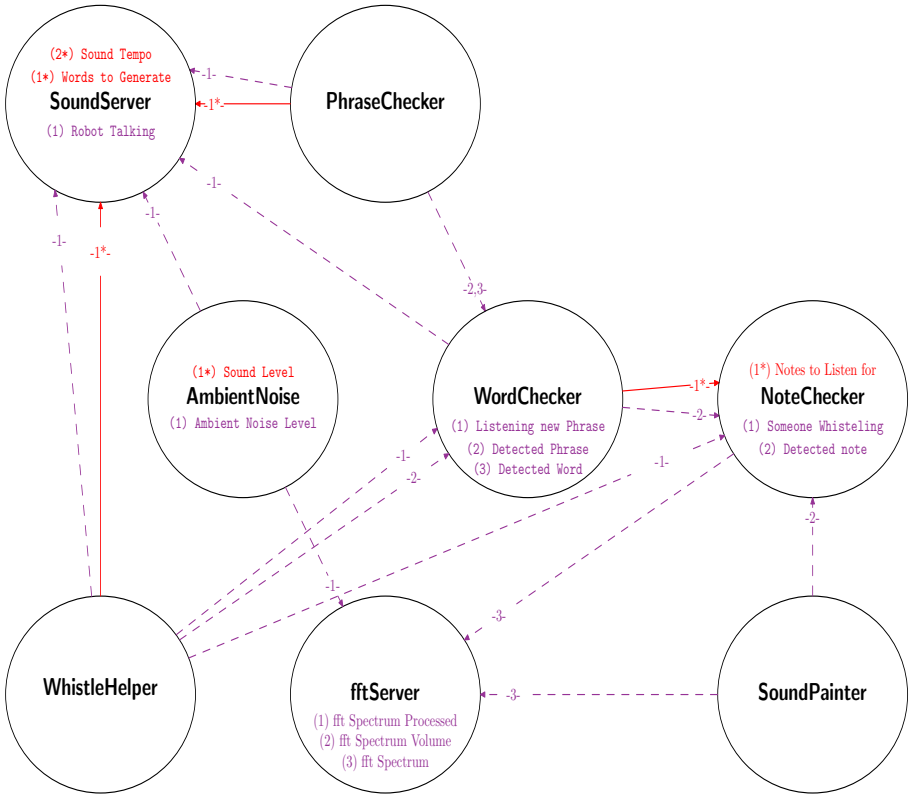
Following Minsky’s concept of *The Society Of Mind*, each agent is designed to perform a specific task, and they are built to exploit the capabilities of other agents to do their own particular tasks. Connecting specialised agents together like this thus forms a kind of society: agents do their own thing, but they depend upon other agents to do it. This results in a widespread reuse of agent capabilities, and also makes it possible to both incrementally develop the agent network, as well as to subsequently extend it to do new things. (It also supports efficient debugging, since errors and failures are usually easy to detect and locate in the code.) Each agent in the society thus adds some particular capability or skill, as Minsky put it.

To perform its skill, an agent typically needs to interact with other agents. It may need data from other agents, and it may also need to instruct other agents about how to do their own work. To support this kind of interaction, two interaction mechanisms have been defined:

1. *Request/Serve Data*: An agent may generate data as output and may require data as input from other agents. Thus each agent must know where to find its own required data. An agent serves data to any agent requesting it, and acts as a client for any data it gets from other agents. When one agent connects as a client to another agent and requests data, the server agent sends them to it. While the connection is open, whenever the data change, the server agent send the new data to all the client agents connected to it. Each agent admits any number of client connections.
2. *Send/Receive Instructions*: An agent may accept certain kinds of instructions from other agents, which may influence the way it performs its task. It may also need to send certain instructions to other agents.

There is thus a possibility for conflicts to arise, since two agents may instruct the same agent to perform its skill in two different ways at the same time. In many cases an agent needs the effect of its instruction to be maintained for a certain amount of time. To deal with this, five priority levels were defined to be assigned to the instructions. Level one is the lowest priority and level five is the highest priority. An agent can lock the priority level of the instruction type accepted by another agent, so that this agent can only accept another instruction from another agent, if it comes with a higher priority





**Fig. 4.** The society of sound agents: a lower-case name *with* a \* indicates an instruction type; a lower-case name *without* a \* indicates output data; continuous arrows represent instruction data; dashed arrows represent served data; and the labels indicate which output data is requested, and which instruction types are sent, or which output data is requested

level, until it is unlocked by the first agent. While a higher level is specified, the instructions with a lower level are ignored. Every time an agent sends an instruction, the instruction receiving agent responds indicating if the instruction has been accepted or not. If it is a priority level lock request, the instruction receiving agent responds indicating if the lock has succeed or not.

## 5.2 The Sound Agent Society

Figure 4 presents the society of agents involved in processing signals from a microphone to acting on interpreted MML phrases, together with their respective data and instruction connections and types. The complete process of recognising MML phrases goes as follows:

1. The **FFTserver** agent generates FFT Spectrum data (Fast Fourier Transform) [18], from digitised samples of the signal from the microphone, and makes it available to any agents requesting it.
2. The **NoteChecker** agent requests the FFT Spectrum data from the **FFTserver** agent, and looks for any MML notes in the sound signal sample.
3. The **WordChecker** agent, requests the recognised notes from the **NoteChecker** agent, and builds them into words and then phrases, based upon the defined grammar of MML. As the recognition is directed to some particular notes, each time the **WordChecker** agent gets a note, it instructs the **NoteChecker** agent about what notes to direct the recognition to. When the **WordChecker** agent detects a whole phrase, it makes it available for any agent that needs it.
4. Finally, the **PhraseChecker** agent gets the phrases from the **WordChecker** agent and, depending on the meaning, sends instructions to other parts of the agent network, not shown here.

The algorithm for detecting MML notes is as follows:

1. Check if the largest amplitude in the FFT data is for a frequency lower than 1046.50Hz, the lowest note on the MML alphabet (see table 2). In other words, check to see if the largest amplitude sound is in the ambient (non-MML) frequency range of the FFT Spectrum data, or in the MML frequency range.
2. If this largest amplitude *is* in the MML frequency range, then for each of the eight notes in the MML alphabet, estimate the power of each note in the signal. This is done by calculating the area under the FFT curve corresponding to the frequency of each MML note  $\pm \frac{1}{4}$  of a whole tone. This results in eight note power values, one for each MML note.
3. For a valid note detection, two further conditions must be satisfied:
  - (a) The largest of these eight power values must be at least three times larger than the third largest power value in the set of eight. (The third largest power value is used for this comparison, because the MML notes immediately above or below the note with the largest power value often have power values close to this largest power value, and so do not help to establish that a MML note has been detected.)
  - (b) The note with this largest power value must be in the list of notes that expected next in the phrase, as given by the instruction type **Notes to Listen for**, from the **WordChecker** agent.
4. A silence is taken to be detected if more than one second elapses in which there has been no successful MML note detections.

The **Robot Talking** data generated by the **SoundServer** agent, indicates when the machine itself is talking. This is used by several agents. One of them is the **WordChecker** agent, which uses it to recognise if a note obtained from the **NoteChecker** agent was generated by the machine or some other source—a person or another machine.

The `WhistleHelper` agent is designed to help people whistle the MML phrases. It gets data from the `NoteChecker` `WordChecker` and `SoundServer` agents, and if it see that the `NoteChecker` agent has detected a note, but that it is not one that is expected by the `WordChecker` agent, it assumes that someone is trying to whistle, but is not whistling recognisable MML words. It then tries to help the person by playing the name of the machine. This both gives the person the correct starting note he or she needs, and helps the person to learn the name of the machine too.

The `AmbientNoise` agent is designed to control the sound generation power (volume) depending on the ambient noise level. It makes the machine speak louder when the general ambient noise level is high, and to speak more quietly when there is little or no ambient noise.

The last agent, the `SoundPainter` agent is used to produce a graphical out put of the FFT Spectrum data, together with the sequence of MML notes detected.

## 6 Evaluation: Real World Trials and Results

From the beginning, Project MiReLa has adopted an unusual approach to evaluating its research results: *all* tests and trial take place in some (un-modified) part of the Main Building of the San Sebastián Technology Park. In contrast to most other robotics research, the real world (the Main Building of the San Sebastián Technology Park, in this case) has always been the testing ground for the research, and not some final demonstration forum. Furthermore, no specially prepared laboratory experiments have been used to evaluate the work, as is the case for much reported robotics research. As a consequence, the evaluation of the incremental investigation of intelligent behaviour via the development of the robot MiReLa as a kind of service robot, is based upon direct observations of what works and what doesn't work in real conditions, and upon a knowledge of what causes the robot to fail *and* to succeed in these real world situations. In other words, quantitative data from laboratory experiments are *not* used to claim that the robot works; MiReLa's performance in the real world is.

Given that the Main Building of the San Sebastián Technology Park is used throughout the year to host many conferences, exhibitions, workshops, presentations, and other professional meetings, as well as concerts, receptions, and film production sessions, there are many opportunities to test MiReLa in a wide range of real conditions and situations, including a wide range of different kinds of people. One particular kind event that has been important in this respect, has been the annual San Sebastián Technology Park Open Day.

Each year, in November, an Open Day event is held. The aim of these Open Days is to present to a general public the work of the San Sebastián Technology Park, and the various research centres, companies, and other organisations situated in the Technology Park. To date, there have been five such Open Days, the first in 2000, and on such occasions about 2,000 people visit the main building, during the morning. MiReLa has been presented and demonstrated during all five Open Day events.

On these Open Days, the Main Building, especially the Hall is transformed by many stands and presentations from the various companies in the Technology Park. It also fills up with people and noise. These events have therefore been very important occasions for testing the robot *MiReLa* in some tough but realistic conditions. Many decisions about the directions of the subsequent research have been strongly influenced by the evaluations and outcomes of these Open Day trials.

In this section we first describe the three Open Day events (held in 2002, 2003, and 2004) that have been important in the evaluation of the research and development of *MiReLa Music Language* and the *Societies of Agents* that implements it. Afterwards, we describe several further evaluations of *MiReLa Music Language*.

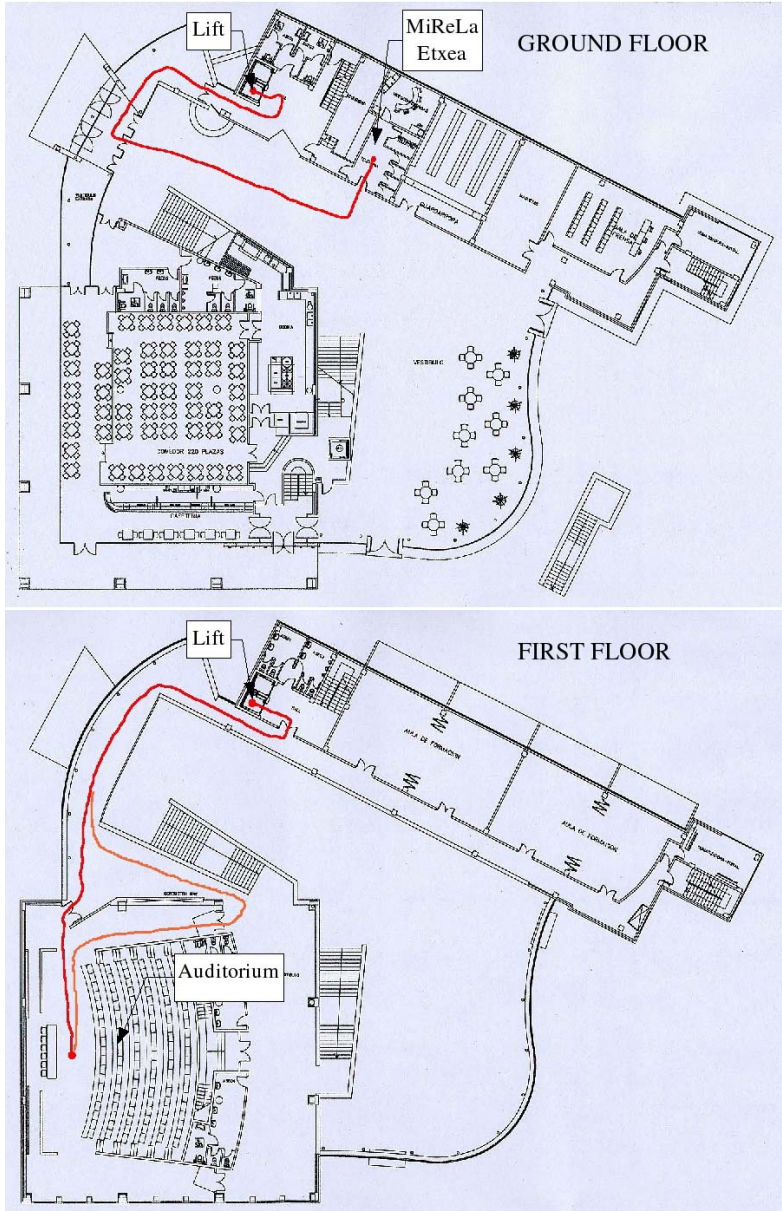
### 6.1 *MiReLa* at the Open Day Trials

**Open Day 2002.** At the 2002 Open Day, *MiReLa* guided a person from “*MiReLa* Etxea” (its home), on the ground floor, to the Auditorium, on the first floor of the building, and then returned home again. Figure 5 shows the plans of the ground floor and first floor of the main building, and the trajectory followed by *MiReLa*. The total distance of the whole tour was more than 200m. To complete it, the robot had to pass through ten doors and to go up and then down one floor in the lift.

A mobile camera crew transmitted a live video stream to the Auditorium, and a Webcam was installed on the robot, so that people seated in the Auditorium could both watch what was going on, and to see a *MiReLa*-view, projected on the large screen in the Auditorium.

Two runs of this trial were presented; one at 11:20 and the other at 13:20. For the first run there were not so many people wandering around the building, for the second run it was much more crowded. On both occasions *MiReLa* successfully navigated its way from its home, through the Hall, round to the lifts, up the lift, and then to the Auditorium and, and successfully returned to its home.

An early version of the *MiReLa Music Language* was tested at this event. At this time, although the robot could generate phrases with name, verb, and arguments, it could only recognise phrases of one word, such as its name (“*MiReLa*”), “hello” and “continue”. This kind of interaction was used with the robot in the lifts scenario, or in front of a door. Recall that *MiReLa* doesn’t have any kind of devices to help it operate the lift, nor to open and close doors. Thus it needed the help of a person to open and close doors and to operate the lift for it, at this time. For the closed door situation, the robot detected when a door was closed and it said “open door”, in MML, so that the person following it could open the door. At the lifts *MiReLa* stopped in front of the two lifts, and waited to hear the word *continue* in *MiReLa Music Language*. The person accompanying the robot called the lift and kept the doors open, then he played the *continue* instruction on a harmonica. The robot went into the lift and said which floor number it wanted, again using *MiReLa Music Language*. It then waited for a *continue* instruction again. The person operated the lift to take it to the required



**Fig. 5.** Plans of the ground floor and first floor of the Main Building of the San Sebastián Technology Park, with the trajectory followed by MiReLa at the 2002 Open Day from MiReLa Etxea, its “home”, in a room just off the main Hall, out of the Hall round to the lifts, up the lift, out and around to the Auditorium on the first floor, and then back to the Hall, leaving the Auditorium by a different exit

floor. When it arrived, the person kept the doors open again to let the robot exit, and then played a *continue* instruction again, and the robot made its way out of the lift.

For the 2002 Open Day trial, no society of agents was implemented. The sound recognition and generation process was formed by two threads, one for doing the recognition from the sound signal to a Fast Fourier Transform Spectrum and then to recognised MML notes and words, and the other to show the Fast Fourier Transform Spectrum values on the screen.

It was observed that most people following the robot during these two runs easily understood when MiReLa was asking for a door to be opened, in front of a closed door. People made comments such as “look it is saying that the door is closed” or “it is saying: open the door!”. These results helped to confirm that MiReLa’s use of MML did help people understand what was going on and what the robot was trying to, even though they did not understand the MML itself. This was important because one of the reasons for developing MML was to try to keep the people MiReLa is supposed to help, as a guide robot, involved in the on-going situation, and in a way that they understand what the robot is doing and why. But without MiReLa using some kind of synthesised speech system.

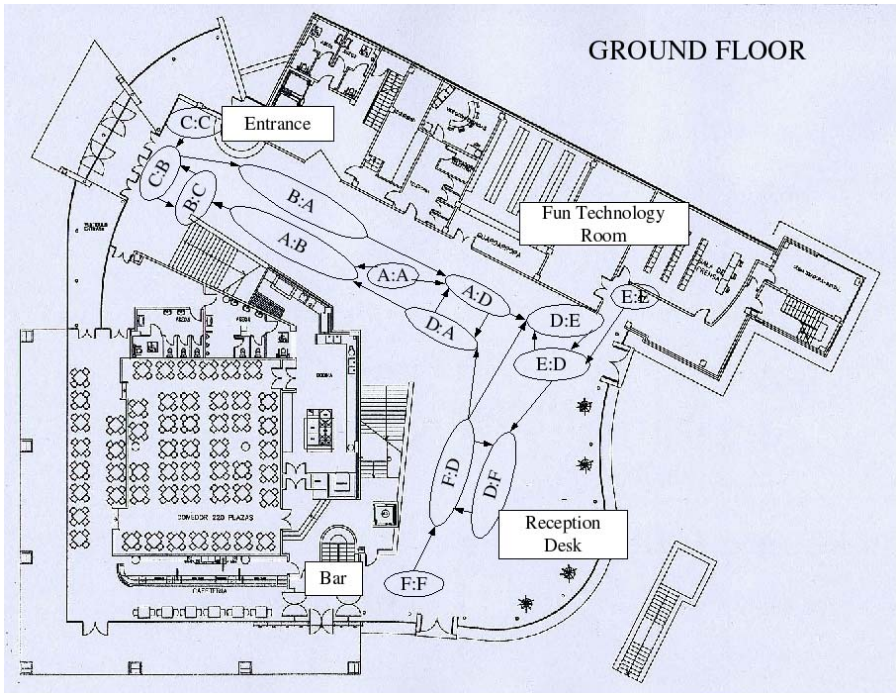
From the outcome of this 2002 Open Day, it was also concluded that the interaction between the robot and the lift should be a *MiReLa Music Language* type of interaction to keep to the style or machine-people interaction being investigated. This way, even if people following the robot were not able to understand the meaning of the sounds generated by the robot and the lift, they would understand that a robot-lift communication was taking place between the two machines. It was also concluded that the signal and MML processing needed to be improved, to make it more reliable and robust, and to improve the structure of it’s implementation.

**Open Day 2003.** At the 2003 Open Day, MiReLa was demonstrated as an aid for the reception service of a large building. It offered a video-conference option to contact a receptionist, and a Web page interface for the receptionist to (remotely) control the robot from his or her desk. (See [11] for more details.)

For this trial MiReLa operated in the Hall of the Main Building, and an early version of the *Society of Agents*, was implemented that integrated the robot navigation system and the MML-based interaction. Figure 6 shows the plan of the ground floor of the main building, with the *Transition Group* graph used by MiReLa to go from one place to another.

The robot started and returned to a place near to the entrance of the main building (labelled “Entrance” in figure 6), doing *Transition Group* C:C, from where it offered the video conference connection to another person placed at the point labelled “Reception Desk” in figure 6. A person arriving at the “Entrance” could then tell the person at the “Reception Desk”, via the video-conference link, where (in the Hall) he or she wanted to be guided to by the robot.

The person at “Reception Desk” then used the Web page interface, to instruct MiReLa to guide the person to his or her desired place. In other words, MiReLa



**Fig. 6.** Plan of the ground floor of the Main Building of the San Sebastián Technology Park, with a representation of the *Transition Group* graph used by MiReLa to navigate to the different places on the ground floor, at the 2003 Open Day

was instructed to be doing either *Transition Group* A:A, *Transition Group* E:E, or *Transition Group* F:F, in figure 6.

In this trial, MiReLa operated continuously from 10:00 to 14:00, and it always reached the places it was asked to go to. At around 12:00, the Hall became so crowded that it was necessary to slow down MiReLa's speed to prevent it from bumping into people in front of it, because they did not have enough room to make way for it. Figure 7 shows two images taken of the occupancy level of the Hall between 12:00 and 13:30.

The society of agents, made it possible to split what was done by a single process before into several tasks done by different agents. It gave flexibility and robustness to the software implementation as it could be developed incrementally, simplifying also failure detection. In this early version of the society of agents, there was not a clear distinction made between instruction and request/send data communication and the priority based instruction system was not implemented yet.

Sound recognition and generation was split into a set of agents consisting of: FFTserver, SoundServer, NoteChecker, WordChecker, and PhraseChecker. For this Open Day trial, the robot could be commanded in real time to go



**Fig. 7.** Pictures showing the occupancy level of the Hall of the Main Building of the San Sebastián Technology Park between 12:00 and 13:30 during the 2003 Open Day

to different places, and a more elaborated version of the MML was needed. In other words, phrases composed of a verb and argument(s) were required to be recognised by the robot. It was decided to place first any arguments a verb may have and terminate the phrase with the verb. When a verb was recognised, the phrase was finished, thus any words recognised before were treated as the verb's argument(s). If more than one argument was recognised and the verb was a one argument verb, the argument last heard was treated as the valid one for the verb. This was thought to be a good way to handle verbs with variable number of arguments and a clear way to separate phrases.

The development and preparation of this 2003 Open Day demonstration showed, however, that for the tasks the robot is able to do, only verbs with one argument or no argument are needed. As a result the definition of MML was later modified to the version presented in section 4.

For sound recognition, there was no directed listening implemented (see section 4.3). In other words, the `WordChecker` agent didn't instruct the `NoteChecker` agent about which notes to try to find in the FFT data. The algorithm for note detection was also different from the one explained in section 5.2: a fixed threshold value was used for the amplitudes of the input sound frequencies served by the `FFTserver` agent. If an amplitude larger than this threshold was detected at some frequencies, the largest among them was chosen. If the same frequency was found in the FFT data arriving during a fixed amount of time, it was concluded that a note of that frequency had been detected.



Using a fixed threshold, required a precise control of the input amplification selected for the microphone used to capture the sound. If it was too high, there were many false detections, and if it was too low, no detections occurred. Ambient noise also had a large influence on the performance of the note detection algorithm. If it was high, there were many false detections (notes detected from the ambient noise). The fact that the phrases started with an argument increased the probability of recognising a valid word from the false detections, since a verb can have many different arguments.

For each place the robot was able to navigate to, a different noun formed by three notes was created. With the robot able to navigate to many different places and starting the phrases with the arguments, almost any three notes could form an argument. In places with high ambient noise, arguments were detected from the ambient noise, and sometimes even a whole MML phrase. This experience served to enforce the determination to continue investigating the use of the music language approach, and avoid any attempt to use any Human Natural Language, since it would not be very unlikely to achieve an unreliable speech recognition in such conditions. But it was from this Open Day trial that it became clear that a more reliable MML system was needed.

**Open Day 2004.** For the 2004 Open Day a better visitor-robot interface was developed using a fixed Information Point, from which it was possible to tell MiReLa to guide a person to any part in the ground floor and first floors of the Main Building of the San Sebastián Technology Park. Also the system described in section 3, to operate the lift using DTMF sounds, was installed and tested.

MiReLa operated from 10:00 to 14:00 and it was able to go to any place on the ground floor and the first floor of the Main Building of the San Sebastián Technology Park.

Interaction with MiReLa was via the Information Point and using *MiReLa Music Language*. Based on the results of the 2003 Open Day, directed listening was developed (see section 4.3) and the sound recognition algorithm explained in section 5.2 was implemented in the robot. This algorithm removed the dependence on a threshold value for note detection, and so the amplification level for sound acquisition, or the ambient noise level didn't influence in the note detection. It also removed the need for defining a fixed amount of time to determine that a valid note had been detected. The syntax of the musical language was also reformed. (But it didn't yet have the form of the syntax presented in section 4.)

In order to use directed listening, a better option was to change the order of the elements of the phrases. The verb now preceded its arguments so that after a verb was detected the listening could be directed only to the notes that form its possible argument words. The decision to start any phrase directed to the robot with its name was also implemented for this Open Day event. Starting the MML phrases with the name of the robot, thus acted as an effective filter against detecting valid phrases in *MiReLa Music Language* from any background noise. The directed detection used in this trial, also reduced the possibility of the robot making false detections in subsequent parts of any MML phrases it heard.

The improvement achieved in the sound recognition system was significant. This time the system worked reliably even between 12:00 and 13:30, when the building was very full of people: MiReLa didn't receive any false positive detections.

For the 2004 Open Day, the system to operate the lifts using DTMF sounds was installed for one lift. As a result, the robot is able to operate the lift itself. The spaces in front of the lifts, on each floor, have rather poor acoustics. This is due to their geometry and surface material, and so any sounds generate a large echo. During testing of the robot for the 2004 Open Day, it was noticed that in this environment, the *MiReLa Music Language* recognition was not as reliable as desired for recognising words containing consecutive repeated notes (for example, *ccd*). From these results a new reform of the music language was designed, which resulted in the version of the MML explained in section 4.

Also more agents were developed and used for this Open Day event. The priority based instruction system (see section 5.1) was not yet implemented and some problems started to arise as the probability of two agents instructing the same agent at the same time increased. From these result, the need to handle conflicts was identified.

## 6.2 Further Trials and Test

On 23 June, 2005, the robot MiReLa and the lamp MiFaRe were transported to the Moncloa Palace (the official residence of the President of Spain), in Madrid. This was for a presentation of a new National research funding programmed, INGENIO 2010, [25]. Three hundred people, politicians, business people, and research directors, were invited to the event. After the formal presentation of the INGENIO 2010 programme, the visitors had time to view the different research projects that were presented, including Project MiReLa. During this time, MiReLa operated continuously for two hours and a half.

For this presentation, MiReLa was situated outside, on the veranda at the entrance of the "Consejo de Ministros" building. It had a (very small)  $2m^2$  area to move in. The lamp was placed on one side of the area. MiReLa could be commanded, using *MiReLa Music Language*, to do different things, such as rotate a specified number of degrees, play a song, or move to the lamp, or away from it. The sound detection system presented in this article was implemented for this test. When the robot moved to the lamp, it switched it on or off (depending on the state of the lamp), by playing the notes of the name of the lamp (*efd, MiFaRe*).

The Moncloa Palace was a completely unknown and unprepared environment for MiReLa, so it constituted a new kind of test. At the end of the presentation, MiReLa also wandered around part of the ground floor of the "Consejo de Ministros" building, avoiding obstacles and people around it. It never bumped into anything, or anyone. This was a motion control test of MiReLa in a new and quite different place.

Apart from the ambient noise generated by the three hundred guests, there was music being played in the background. This was not a handicap for MiReLa and the lamp, MiFaRe, to reliably understand the more than forty commands

given to them from different music sources, such as a xylophone, a harmonica, a mobile phone, or by people whistling. The lamp was switched on and off by MiReLa more than fifteen times and it never failed in any of these attempts. The priority based instruction system (see section 5.1) was implemented for this trial.

Tests have also been carried out with the robot MiReLa and the lamp MiFaRe, to establish the distance, from someone whistling MML commands, that they can reliably recognise MML phrases.

For the robot MiReLa, these test were conducted in the Main Building of the San Sebastián Technology Park, with a person standing on the balcony overlooking the main space of the Hall. The usual background music (“musik”) was also being played at the time. Under these conditions, MiReLa was reliably able to detect and correctly recognise MML phrases whistled to it from up to fifteen meters.

In the case of the lamp, MiFaRe, a person whistled the lamp’s name from just outside the office the lamp is located in, while music, by the group “Rage against the machine”, was played at a loud volume—more than is normal when people are working in the office. Once again, there was no problem to make the lamp switch on and off using the *MiReLa Music Language* under these conditions.

Since the 2004 Open Day several tests have been carried out to check the difficulty people have to whistle words in the *MiReLa Music Language*. These tests consisted of trying to encourage people, on seeing the robot MiReLa or the lamp MiFaRe for the first time, to whistle the name of the machine.

So far, twenty-one people have tried whistling to the robot MiReLa or to the lamp MiFaRe, and the *WhistleHelper* agent on MiReLa was used to help them to do this. Eighteen people succeed in their attempts. Nine of the people who whistled to the machines were children, and they all succeeded in getting the machine to detect their respective names. During the event INGENIO 2010 at the Moncloa Palace, eight people whistled to the lamp MiFaRe or to the robot MiReLa. Six of them succeeded in their attempts.

Some people could whistle the name at the first attempt after listening to the notes given by the *WhistleHelper* agent. Most of them could do it at the second or the third try.

## 7 Discussion and Conclusions

The development and use of *MiReLa Music Language* goes against the general current of work in human-machine interaction today. Most other work, especially in intelligent robotics, try to make robots that look and behave like humans; by giving them faces, Human Natural Language interfaces, and even humanoid forms. The theory, or perhaps better said, the presumption behind these attempts is the idea that to make machines easier for us to use, they need to be more like us. And the implication is that the more like us our machines are, the easier they will be to use. Project MiReLa, and, in particular, the development of *MiReLa Music Language*, both rejects this theory and questions its validity.

Human-like faces and human-like forms are *not* functionally necessary to have useful and effective robots. They therefore introduce unnecessary complexity and cost to the development, construction, and control of robots and other machines. The full expressiveness, subtleness, openness, and complexity of Human Natural languages is also far beyond anything we need to communicate effectively with the kinds of machines and robots we have today, and will have in the (real) foreseeable future. And, as was commented upon the introduction, the poor approximations of Human Natural language using systems that we have today, require us Humans to adapt and distort the way we use our own languages, just so that the machines can understand us.

Even supposing these technical difficulties can be solved, complexities can be resolved, and costs can be reduced by future advances, we are still left with what we believe is the more important question behind all this: do we really want our machines and robots to be made more and more like us? We believe not, and our experience in developing and testing *MiReLa Music Language* with MiReLa, the robot, and with MiFaRe, the desk lamp with real people in real conditions have encouraged us to continue to believe this. Furthermore, the reliable and robust human-machine interaction we have achieved with MML in real-world conditions also serve to show that the approach we present here is a practical and realistic alternative, and one that people see to like and be happy with.

Unlike other sound-based or music-based interfaces, such as the earcons described in section 2, MML has proved to be quite easy to learn, recognise, and use, even by people not directly involved in Project MireLa. This is the result of developing a version of MML that is well designed for whistling, and which gives rise to simple but musical tunes. The fact the people have to use (produce) MML, and not just listen to machines producing it, mean that they learn the language in a way similar to the way we learn our own Human Natural languages: the experience and knowledge of the language gained from having to produce it, helps in listening to it and understanding it. The simple musicality of MML tunes also mean that after a little practice, it is possible to understand MML phrases at the musical (tune) level, and not have to interpret the tunes into MML notes and words. This, again, is not unlike how we understand a lot of what we hear in our own languages, especially in noisy uncontrolled conditions, when we often do not hear all the syllables and words of what is being spoken.

So, although *MiReLa Music Language* is an artificial language, designed for the kinds of simple human-machine and machine-machine communication we need, it does share some important characteristics of Human Natural languages.

On the practical side, the software agent system used to implement MML processing, is computationally small and inexpensive. Together with the relatively simple hardware needed to implement a MML generator and recogniser, a complete installation, in the robot MiReLa, or lifts, or even for an individual desk lamp, and other devices and machines, requires very little resources. Nor is it difficult to install, maintain, or to modify and extend. And, in all three example installations described in section 3, the MML and its implementation

has been shown to work well in real conditions, and to be quite easy and comfortable to use.

These results encourage us to believe that MiReLa Music Language offers a simpler yet natural way for communicating with machines, and therefore to propose it as the basis for a more generally applicable human-machine communication language in ambient intelligence installations for everyday life.

## Acknowledgements

Project MiReLa has been supported by the San Sebastián Technology Park, and many people have worked on the project at different times: we thank them all for their contributions. We are particularly grateful to Manuel Cendoya and to Amaia Bernaras who co-founded the Project MiReLa (with Tim Smithers), and who have provided much essential help and support over the years. We also acknowledge the involvement of Rod Brooks and Leslie Kaelbling, from the Laboratory for Computer Science and Artificial Intelligence, MIT, in various aspects of the development of the project. And, we thank all the people who, often unwittingly, have been a part of the many real-world tests that have been an essential part of Project MiReLa.

## References

- [1] Allen, J.: How Do Human Process and Recognize Speech. *IEEE Transactions on Speech and Audio Processing ASP*, **2**, pp. 567–577, October, 1994.
- [2] Alonso, J. M.: TCP/IP. Programación de las aplicaciones distribuidas. ra-ma, 1998.
- [3] Alty, J. L.: Can we use music in computer-human communication. In *Proceedings of HCT'95*, pp. 409–423.
- [4] American Acoustic Society octave designation system:  
<http://www.music.vt.edu/musicdictionary/texto/Octavedesignation.html>
- [5] Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M.: Earcons and icons: Their structure and common design principles, *Human-Computer Interaction*, 1989.
- [6] Bohlen, M.: Universal Whistling Machine.  
[http://www.realtechsupport.org/new\\_works/uwm.html](http://www.realtechsupport.org/new_works/uwm.html)  
See also, Bohlen, M. and Rinker, J. T. Concurso Internacional sobre Arte y Vida Artificial, U.W.M (2004),  
<http://www.fundacion.telefonica.com/at/vida/paginas/v7/whistling.html>
- [7] Brigham, E. O.: The fast Fourier transform and its applications. Prentice-Hall, Inc., 1988.
- [8] Brewster, S. A.: Providing a structured method for integrating non-speech audio into human-computer interfaces. PhD Thesis, University of York, UK, 1994.
- [9] Brewster, S.A., Wright, P.C. and Edwards, A.D.N.: An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of InterCHI'93 (Amsterdam)* ACM Press, Addison-Wesley, pp. 222–227, 1993.
- [10] Cutnell, J. D. and Johnson K. W.: *Physics*. Wiley, New York, 6 edition, 2004.
- [11] Esnaola U., Rañó, I. and Smithers, T.: Robot MiReLa: Helping the Secretary to Attend Visitors, in *Proceedings of 2nd International Workshop on Advances in Service Robots (ASER 2004)*, Fekdafing, Lake Starnberg, Germany, May 20–21, 2004.

- [12] Esnaola, U., Rañó, I. and Smithers, T.: A Behaviour-Based Navigation System, in proceedings of TAROS'2004, University of Essex, England, September, 2004.
- [13] Esnaola, U. and Smithers, T.: MiReLa: A Musical Robot, accepted for 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA'2005), Helsinki University of Technology, Espoo, Finland, from June 27–30, 2005.
- [14] Gray, W. D.: VCR-as-paradigm: A study and taxonomy of errors in an interactive task. In Nordby K., Helmersen P., Gilmore D. J., and Arnesen S. A. (Eds.). *Human-Computer Interaction—Interact'95*. pp. 265–270, New York: Chapman & Hall, 1995.
- [15] Gomera: see Whistled Speech. Wikipedia, <http://wikimediafoundation.org/wiki/WhistledSpeech/>
- [16] Husband, G.: What's in your Music. [http://www.tnt-audio.com/topics/frequency\\_e.html](http://www.tnt-audio.com/topics/frequency_e.html), 1999.
- [17] Karolyi, O.: *Introducing Music*. Penguin Books, 1965.
- [18] Karu, Z.: *Signals and systems. Made Ridiculously simple*. ZiZi Press, 1995.
- [19] Leplatre, G., and Brewster, S. A.: Perspectives on the Design of Musical Auditory Interfaces. Focus conference on Anticipation, Cognition and Music. Liege, Belgium, Aug. 1998. In Dubois D. (Ed) *International Journal of Computing Anticipatory Systems*, Feb. 1999.
- [20] MICROPIK: PS6E8S, manual de instrucciones. Profesor Blanco 18 Bajos, 46014-Valencia, Spain, 2003.
- [21] Minsky, M.: *The Society of Mind*, Wiliam Heineman Ltd, 1987.
- [22] NORCOMM: Model NC400 Touch-tone ani/alarm encoder instruction manual. 15385 Carrie Drive, Grass Valley, CA 95949, 2004.
- [23] Pierce, J. R.: *The Science of Musical Sound*. Scientific American Books. W. H. Freeman and Company, New York, 1983.
- [24] Rañó, I.: *Investigación de una arquitectura basada en el comportamiento para robots autónomos en entornos semi-estructurados*, PhD Thesis, Facultad de Informática, Universidad del país Vasco (UPV), 2004.
- [25] Rosa M. Tristan: *Zapatero se compromete a duplicar la inversión en I+D+i para el año 2010*. El Mundo, June, 2005.
- [26] Smoloiar, W.: The role of music in multimediam, *IEEE Multimedia*, pp. 9–11, 1994.
- [27] Vicker, P. and Alty, J. L.: Siren Songs and Swan Songs, Debugging with Music. *Communications of the ACM*, **46**, 7, pp. 87–92, July, 2003.

# Speaker Identification and Speech Recognition Using Phased Arrays

Roger Xu<sup>1</sup>, Gang Mei<sup>1</sup>, ZuBing Ren<sup>1</sup>, Chiman Kwan<sup>1</sup>,  
Julien Aube<sup>1</sup>, Cedrick Rochet<sup>1</sup>, and Vincent Stanford<sup>2</sup>

<sup>1</sup> Intelligent Automation, Inc.,  
7519 Standish Place, Rockville, Maryland USA  
([www.I-A-I.com](http://www.I-A-I.com))

<sup>2</sup> The National Institute of Standards and Technology  
100 Bureau Drive, Gaithersburg, Maryland USA  
([www.nist.gov](http://www.nist.gov))  
[vincent.stanford@nist.gov](mailto:vincent.stanford@nist.gov)

**Abstract.** We summarize our research results on an innovative approach to making smart meeting rooms accessible to hands-free users. Specifically, we developed an autodirective system to acquire speech in a noisy room using a microphone array, and to identify the speech from a privileged speaker among others in real time. We successfully established that a commercial speaker-dependent speech recognition product could recognize beamformed speech acquired using our autodirective algorithm. We used the NIST Smart Flow System and the Mk-III microphone array developed by the National Institute of Standards and Technology to conduct our experiments.

## 1 Introduction

The NIST Smart Data Flow System provides a platform to develop standards that promote the interoperability of disparate sensors and devices produced by different manufacturers. The NIST system can acquire and process multiple sensor data streams, such as voice and image, in real-time. Applications such as speaker verification, head trackers, and the like, can be implemented under the framework. Here we use it to construct a proof-of-concept of the *user sensitive interface* proposed by Stanford in [1] and discussed further by Flanagan and Stanford in [2]. Fundamentally, the proposal was to identify individual users among a working group, such as a situation room or command center, using multimodal sensor data from microphone and camera arrays. We believe that this is a foundational element in effective context-aware systems that can respond appropriately to individuals generally; and it may be especially helpful to those with special needs.

We developed and demonstrated a real-time speaker verification and speech recognition subsystem using the smart data flow system for data transport and real time parallel processing. Speaker verification techniques are widely used for access to private information, personal transactions, and provide security in computer and communication networks [3] and so offer an important research topic. In this application, we use the technique to identify a privileged speaker in a smart meeting room among a group of people and decode his or her speech.

The NIST Mk-III microphone array, which is part of the Smart Space system, provides data to locate each individual speaker in the smart meeting room using a steered response beamformer. Once acquired, the speech is processed by the speaker verification algorithm and forwarded via a data flow to the IBM ViaVoice dictation system for speaker dependent speech recognition. The speaker verification algorithm that we implemented is based on Gaussian Mixture Models (GMM's) of cepstral feature vectors [4]. This algorithm is well studied and can be implemented on modern microprocessors. However, making it work well requires careful selection of model parameters, and interfacing multiple algorithms. The NIST Smart Data Flow System plays a key role as an integrating platform, and properly fine-tuning the GMM, and our Speech Activity Detector was necessary. We found that speech identification works better with a different activity detector from the one optimized for speech recognition. The final demonstration showed successful deployment of our real time speaker verification system in a small group.

Section 2 below gives an overview of the smart flow system. Section 3 details the implementation of our speaker verification algorithm, and its integration using the smart flow system with the NIST Mk-III Array. Results from the real-time demonstration are given in Section 4. We conclude with list future research and development directions.

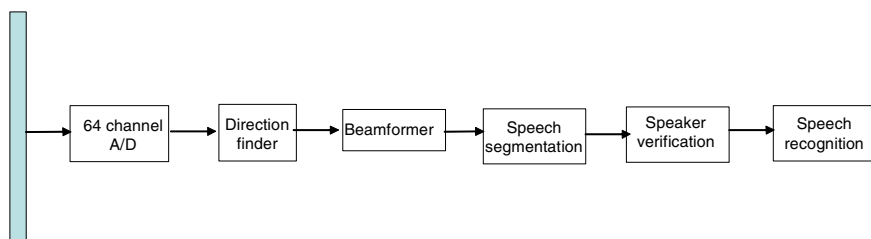
## 2 Overview of the NIST Smart Data Flow System

Figure 1 shows the architecture of our user sensitive hands-free interface demonstration, using the Smart Flow system, which was initiated about five years ago at NIST. After several years of intensive effort, the system has reached to a stage that permits real-time processing of high bandwidth multimodal interface sensor data streams. As can be seen, the system consists of several subprocesses. A linear microphone array with 64 microphone elements is used for speech acquisition and sixty-four-channel analog-to-digital conversion is performed on eight microphone banks with eight microphones each. Our demonstration used a digitization rate of 22kHz for the speech signals because it is compatible with many commercial speech recognizers. The Direction Finder locates the signal sources by converting the array microphones signals into a *beam space* of sixty-four beams and then finding a peak in the steered response curve. The direction information is then used by the autostereodirective speech acquisition system to select a suitable beam that maximizes the desired speaker's voice; and at the same time suppresses the other signal sources.

After beamforming the collected speech is passed through the speech segmentation procedure prior to speech recognition. A threshold is used to determine the presence of speech signals and a state machine using a different trailing threshold and an intra-utterance silence threshold is used to recognize the utterance end-point. The segmented speech signals are further reduced by a threshold procedure that selects voiced speech for verification by a GMM-based classifier to determine the speaker's identity as either the privileged speaker or someone else. After the speaker has been identified, an appropriate speech recognition model for the privileged speaker will be recalled and the speech signal will be recognized for its spoken words. One possible



application would be to display the recognized speech on a screen for hearing impaired, for example; or to parse it for command content and send the resulting commands to software applications of interest to the user.



**Fig. 1.** Architecture of the user sensitive interface as described by process blocks that were mapped to Smart Data Flow System processes. The bar at left represents the NIST Mk-III Microphone array.

There are many potential applications of the Smart Flow system. Here we describe one that could serve people with disabilities, or command center personnel working in small groups. In the U.S., Legislation has affirmed the rights of disabled individuals to a public education, and to physically accessible educational and workplace buildings. However, while an increasing number of buildings are becoming accessible, often the furniture and tools within them are not; this is particularly true of software tools. Technology that is not accessible can impede one's ability to acquire and utilize information. Furthermore, the inability to use technological tools can limit individual's career opportunities, as many of today's jobs, and more jobs of the future, require facility with technological tools. To ameliorate these problems, Section 508 of the Rehabilitation Act was amended by Congress in 1998, and requires that Federal agencies' electronic and information technology be accessible to people with disabilities. In addition, Section 508 mandates that disabled individuals, who are members of the public seeking information or services from a Federal agency, have access to, and use of, information and data available to individuals without disabilities. Section 508 lists technical standards for accessibility to the following: 1) software applications and operating systems; 2) web-based intranet and Internet information and applications; 3) telecommunications products; 4) video and multimedia products; 5) self-contained, commercial off the shelf products; and 6) desktop and portable computers.

Smart space environments have the capability to meet many of these standards and requirements, and they have a strong potential to improve the quality of lives of individuals with disabilities, as well as accelerating the activities of key individuals. In one smart space scenario, an individual in a wheelchair could enter a room, where signals from cameras and microphone arrays could be used to recognize this individual, and automatically open speaker identification, speech recognition, and some common applications. Piping the spoken commands from the privileged user to the applications would offer completely hands-free operation. Other team-based work could be facilitated in new and unique ways as well. For example, one speaker sensitive user interface process could be spawned for each individual in a meeting. The same microphone array data could be used by multiple person recognizers; each

with its own speaker dependent speech recognition program. This would allow workers in command centers, or situation rooms, easy access to information resources via spoken commands.

### 3 The Speaker Verification Subsystem

Our objective was to apply a relatively conventional speaker identification / verification algorithm to the voice data collected by a microphone array using our autodirective speech acquisition algorithm. We found that a different speech activity detection algorithm was appropriate to the speaker identification task than the one normally used for speech recognition. For our demonstration, we reduced the task to speaker verification from the more difficult speaker identification task by classifying a privileged speaker versus a population model for unknown speakers. Thus we need only reject speech segments from other speakers. In addition to the speaker voice, ambient room noise and reverberation of the voice enter the microphones via many reflective acoustic surfaces. These unwanted signals could degrade the performance of the system. Beamforming provides considerable resistance to multipath, reverberation, and extraneous noise sources. Once the Smart Flow system successfully identifies this primary speaker, it triggers other parts of system such as the speech recognition system, and automatic speech-to-text conversion.

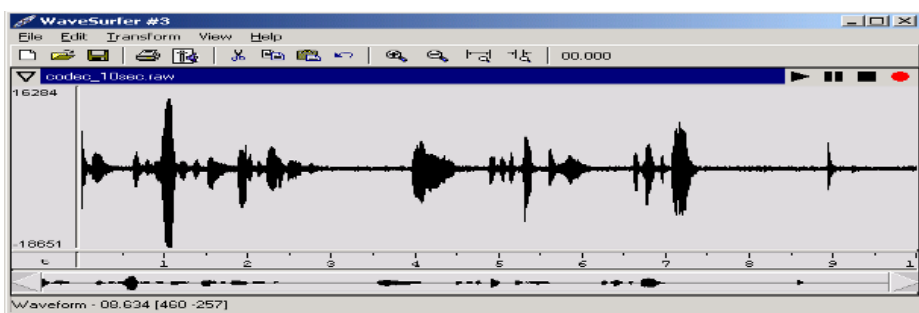
Before we describe the details of the speaker verification system, it is worth mentioning particulars of speech segmentation that are important to the verification subsystem. A correct detection of speech segments is crucial to successful speech recognition, speaker training, and speaker verification. We hence investigated threshold-setting procedures to detect speech activity and provide input speech to the speaker verification system.

First, raw speech data is divided into non-overlapping frames. For each frame, the standard deviation is calculated, and if it is greater than a pre-set threshold, say, 4 times the minimum standard deviation of the last 100 frames, then this frame is considered to be a speech frame. Otherwise, it is considered to be background noise. Thus, two important parameters here are the frame size (in ms) and threshold level (how many times the minimum standard deviation).

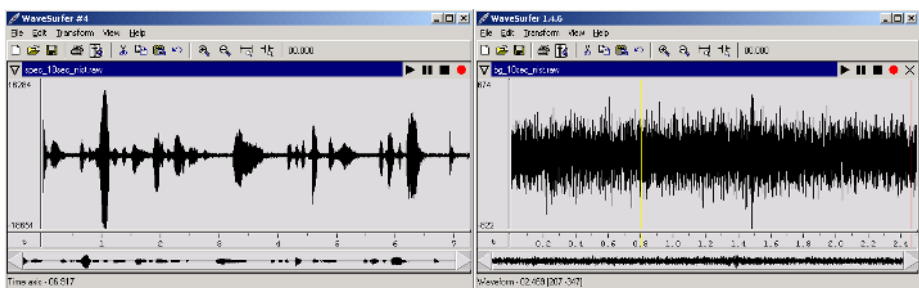
A different algorithm and parameters were implemented by NIST for purposes of speech recognition that used a speech onset threshold with a pre-onset buffer for plosives and fricatives, an intra-utterance silence state machine, and a lower offset threshold for trailing plosives and fricatives. The NIST algorithm was constructed to support speech recognition and hence provides a full continuous segment including possible leading and trailing unvoiced speech. This is particularly important for the first and second difference of the cepstral frame vectors, because the differences are erroneous if individual frames within the utterance segment are dropped. We considered this to be a fairly good choice after many experiments for speech recognition. But we applied a different and simpler speech activity detector for speaker identification and compared the speaker identification system performance with different parameters. The following figures will show the difference between our algorithm and the NIST speech segmentation algorithm.

From these figures we found that, the NIST speech recognition threshold and segmentation algorithm split the 8.634 seconds of raw voice data, shown in Figure 2 into 6.917 seconds of speech, (Figure 3-a) and 2.469 seconds of noise (Figure 3-b). Using the IAI threshold parameters and algorithm, it was split into 2.108 seconds of speech (Figure 4-a) and 6.449 seconds of noise, and unvoiced speech (Figure 4-b). Thus, we basically excluded some speech data, particularly unvoiced speech, as noise but the speech data we keep is shown to be better suited for speaker identification.

We believe that we are thus relying on voiced speech as the primary source of discriminating feature vectors and excluding unvoiced speech and background noise. The speech recognition thresholds must preserve unvoiced speech made up of fricatives and plosives for accurate word recognition, but these are less effective in speaker identification. Further detailed research is needed to rigorously quantify this interesting finding.



**Fig. 2.** Raw voice signal for about eight and a half seconds of speech mixed with silence used for speech segmentation experiments



(a) 6.917 seconds of speech;

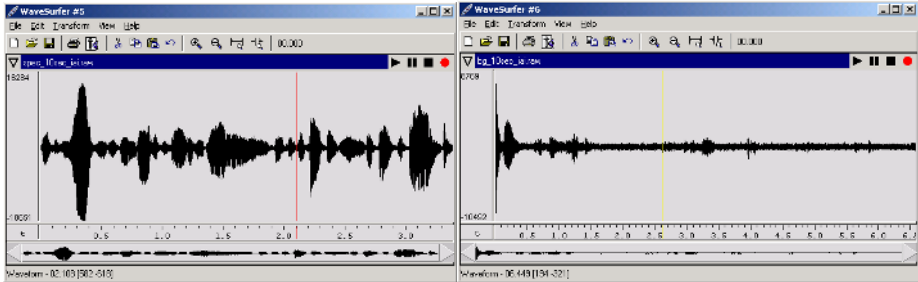
(b) 2.469 seconds of noise

**Fig. 3.** Segmentation resulting from NIST speech recognition thresholds. These settings are designed for detection of and inclusion of leading and trailing fricatives and plosives for speech recognition. Some intra utterance silence remains in the segment, and almost all excluded signal is true background noise.

### 3.1 Preprocessing to Extract Features for Speaker Recognition

To identify the primary speaker, our algorithm first extracts feature vectors from the threshold-selected voice data, then matches these feature vectors with two Gaussian Mixture Models: the primary speaker GMM, and the background population model

GMM. The difference between the likelihoods of the given speech segment under each model is compared to a pre-set threshold to decide if the speech is that of the primary speaker. Figure 5 describes the data flow for the speaker verification subsystem.



(a) 2.108 seconds of speech;

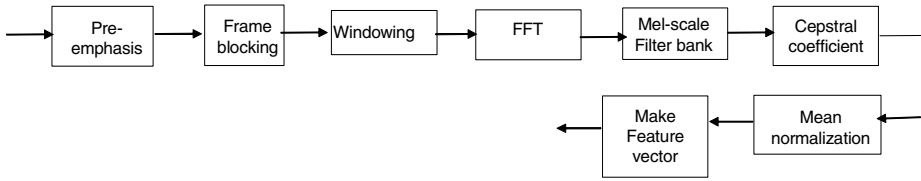
(b) 6.449 seconds of noise

**Fig. 4.** Segmentation resulting from threshold settings designed for speaker identification. A more aggressive and frame-by-frame noise rejection algorithm results in more background noise rejection at the clear cost of rejecting some unvoiced speech frames along with the background noise.

This Mel-scale cepstral feature extraction pipeline is now quite standard and is in wide use in the speech recognition community. It is well documented in textbooks such as [14], and [15] and many other sources in the speech recognition literature.

Our system uses cepstral coefficients derived from a Mel-frequency filter bank to represent short-term speech spectra. The digital speech data is first preprocessed (pre-emphasized, set to overlapped frames and Hamming windowed) and then a Mel frequency filter bank is applied to the cepstral coefficients computed using an FFT. Typically this feature extraction process compresses around 256 samples of speech time series to between perhaps thirteen to forty features depending on the exact configuration of the filter banks and if first and second difference of the base features are used.

In our speaker identification algorithm, the feature vectors are generated in the same way as the feature vectors for speech recognition. However, different sets of parameters are selected to fit the characteristics of the speaker identification algorithm. The preprocessing involves eight steps: a) Pre-emphasizing the signal; b) Frame blocking; c) Windowing; d) Discrete Fourier Transform; e) Mel-scale filter bank; f) Mel frequency cepstral coefficients (MFCCs); g) Adaptive mean normalization; and h) Formation of the feature vectors. The last step of the preprocessing is to form a feature vector by combining the output vector of mean normalization and its 2<sup>nd</sup> degree derivative vector. The feature vector will be used to train and verify speakers. As a result, we obtained a 38 dimensional feature vector by combining a 19 dimension mean normalized vector and its 2<sup>nd</sup> degree derivative vector. Details of each step used in our system can be found in [5] and at ([www.nist.gov/smartspace](http://www.nist.gov/smartspace)).



**Fig. 5.** Feature extraction steps in the speaker-verification subsystem implement a Mel-frequency cepstrum feature vector computation

Using linear microphone arrays, here the NIST Mk-III array, for beamforming can provide bearing estimates for speech sources, which can then be acquired at good signal-to-noise levels at distances commonly found in conference rooms. Earlier work with autodirective speech acquisition systems and beamforming can be found in [6], [7], [8], and [9]. Excellent research approaches to multi modal interfaces being pursued in current research can also be found in [10], [11], [12], and [13].

### 3.2 Gaussian Mixture Models for Speakers

The Gaussian mixture speaker model is a probabilistic model by which the distribution of the data is modeled as a linear combination of several multivariate Gaussian densities. There are two motivations for using Gaussian Mixture Densities to represent speaker identity. The first is the intuitive notion that the individual component densities of a multi-modal density, like the GMM, may model some underlying set of acoustic classes. The second motivation is the observation that a linear combination of Gaussian basis functions each centered at a data point (a generalization of the Parzen Window concept for the estimation of density functions [15]) is capable of approximating a large class of sample distributions. The GMM shares this property, and is trained with the Expectation-Maximization (EM) algorithm to maximize the likelihood of the observed data given the model parameters.

A Gaussian mixture density is a weighted sum of  $M$  component densities, given by the equation:

$$p(X | M, \Sigma) = \sum_{i=1}^M P_i \cdot p(X | \mu_i, \sigma_i) \tag{1}$$

Where  $X$  is a  $D$ -dimensional random vector,  $p(X | \mu_i, \sigma_i)$  are the component densities, and the  $P_i$ , are the mixture weights. Each component density is a  $D$ -variant Gaussian function of the form:

$$p(X | \mu_i, \sigma_i) = [(2\pi)^{D/2} \cdot |\sigma_i|^{1/2}] \cdot e^{-(X - \mu_i)^T \sigma_i^{-1} (X - \mu_i)} \tag{2}$$

Where  $\mu_i$  are the mean vectors; and  $\sigma_i$  are the covariance matrices. The mixture weights satisfy the constraint that  $P_1 + \dots + P_M = 1$ , and  $P_i \geq 0$ . The mean vectors, covariance matrices, and mixture weights from all component densities parameterize the complete Gaussian mixture density. These parameters, for a given speaker  $s$  are collectively represented using the notation:

$$\lambda_s = \{P_{i,s}, \mu_{i,s}, \sigma_{i,s}\} \tag{3}$$

For speaker identification, his or her model represents each speaker  $s$  as  $\lambda_s$ . The GMM can have several different forms depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component (nodal covariance), one covariance matrix for all Gaussian components in a speaker model (grand covariance), or a single covariance matrix shared by all speaker models (global covariance). The covariance can also be full, diagonal, or even circular.

We used a well-documented Matlab toolbox called Netlab to perform the GMM model estimation and other tasks needed in pattern classification. The toolbox, developed by Ian T. Nabney at Aston University in the U.K., provides many useful Matlab functions for speech processing [17].

Our implementation of the speaker identification algorithm, in addition to creating a GMM model for the primary speaker as described above, uses a GMM to represent the population of unknown speakers. The way of doing this is extremely simple: we combined the cepstral data from multiple speakers and combined their normalized feature vectors and then used them to train a single GMM model as if they all came from a single speaker. We chose the model to have 16 components, and a full covariance matrix for each Gaussian kernel as its statistics.

### 3.3 Speaker Identification from Beamformed Speech

For speaker identification, a group of  $S$  speakers is represented by the speaker specific GMM's. The objective is to find the speaker model, which has the maximum *a posteriori* probability given a particular observed sequence [2]. That is:

$$\hat{P} = \max_{1 \leq s \leq S} P(\lambda_s | \{X_{t=1}^T\}) = \arg \max_{1 \leq s \leq S} \frac{p(\{X_{t=1}^T\} | \lambda_s)}{p(\{X_{t=1}^T\})} \tag{4}$$

The right hand expression is due to Bayes' Rule. Assuming equally likely speakers and noting that  $p(\{X_{t=1}^T\})$  is the same for all speaker models, the classification rule simplifies to:

$$\hat{s} = \arg \max_{1 \leq s \leq S} p(\{X_{t=1}^T\} | \lambda_s) \tag{5}$$

Transforming to logarithms and assuming independence between observations, the speaker identification system uses the computing formula as:

$$\hat{s} = \arg \max_{1 \leq s \leq S} \sum_{t=1}^T \log p(X_t | \lambda_s) \tag{6}$$

In our speaker identification algorithm, we only have two GMM models: primary model and population model. Thus our decision algorithm was simplified as follows:

$$\frac{1}{T} \sum_{t=1}^T \log p(X_t | \lambda_{speaker}) - \frac{1}{T} \sum_{t=1}^T \log p(X_t | \lambda_{population}) > threshold \tag{7}$$

This condition implies that the speaker is the primary speaker rather; otherwise he or she is classified as Unknown Speaker. In this equation,  $T$  is the number of feature vectors needed for speaker identification, and *threshold* is a pre-set value.

In our experiment, using the primary model and Unknown Speaker population model, the correct identification rate with only 100 vectors (each vector corresponds to 256 sample points, which is  $256/22.05 = 11.4$  ms) is more than 90%. This speaker identification algorithm also successfully worked in a real-time system that continues to collect the voice data in an office environment. The system can recognize the primary speaker among other speakers as long as they are not speaking at the same time, in which case there is a high probability of an Unknown Speaker classification.

## 4 Simulation and Experimental Results

For this experiment we used 5977 vectors (each vector corresponds to 256 sample points, which is  $256/22.05 = 11.4$  ms) from the primary speaker to create the primary GMM model and we used 5977 vectors from 4 other speakers to create the population model. The models were trained using Netlab package we described above. The corresponding training time is  $5977 \times 11.4$  ms = 68.1s, which is the pure speech time after applying our selection threshold. Also, we excluded the first five hundred feature vectors to assure that the adaptive cepstral mean normalization had converged.

**Table 1.** The probabilities of missed detection and false alarm with the detection threshold set to zero. The use of voiced speech for speaker classification is shown to be effective for utterance length sequences of cepstral vectors.

| Vectors used | Probability of missed detection | Probability of false alarms |
|--------------|---------------------------------|-----------------------------|
| 1            | 24.4%                           | 24.4%                       |
| 2            | 21.4%                           | 19.6%                       |
| 4            | 15.4%                           | 14.9%                       |
| 10           | 7.6%                            | 8.0%                        |
| 20           | 5.4%                            | 3.7%                        |
| 40           | 2.2%                            | 1.0%                        |
| 100          | 0                               | 0                           |

Table 1 shows that the probability of false alarms and missed detections drops with increasing numbers of feature vectors in the likelihood computation. Moreover, these drop rapidly with increasing numbers of feature vectors upon which to base the classification. Additional models, say Unknown Male, and Unknown Female might improve the performance of the classifier as more speakers are added to the background models. To investigate the speaker identification performance, we tested the two types of errors: the probability of missed detection, in which the primary speaker is identified as a non-primary speaker; and the probability of false alarm, in which a non-primary speaker is identified as the primary speaker.

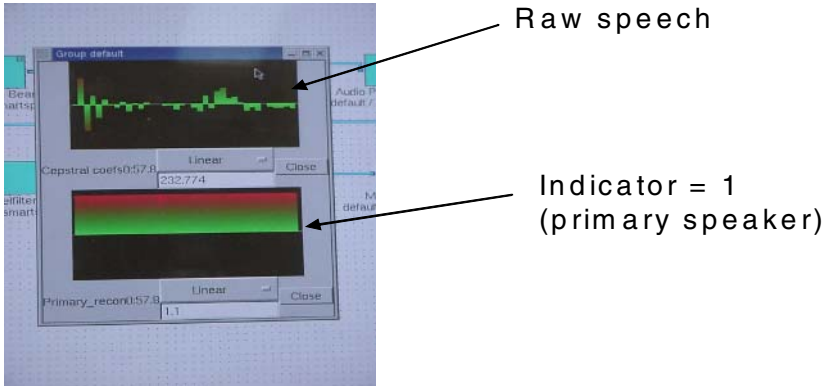
### 4.1 Real-Time Speaker Verification

Our real time speaker verification algorithm accesses its input from a data-flow rather than a batch input file. Thus, in the real-time system, speaker verification is a client

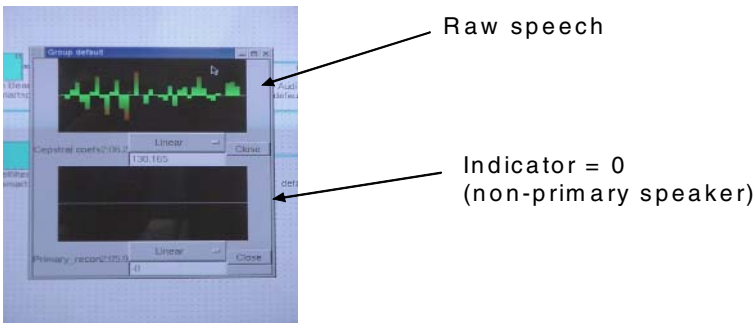
of NIST Smart Data Flow System. This client subscribes to the “Cepstal Coef” data-flow as the source of its input, and opens a “Primay\_recon” data-flow as its output. When the client is running, it buffers the input feature vectors from “Cepstal Coef” data-flow. Once the buffer is full, it scores the data against the two GMM models (Primary GMM model and Population GMM model) and makes the decision. If it decides the data is from primary speaker, it emits “1” to “Primary\_recon” data-flow, otherwise, it emits “0”.

Figures 6 and 7 show the system in operation. The display window for “Cepstral coefs” and “Primary\_recon” data-flows show the primary speaker indicator over a time window. Figure 6 shows the primary speaker classification for the real time speech data stream. Fig. 7 shows the configuration when the real time speech stream classification is non-primary speaker speech. This test was based on speech acquired using the steered response beamformer that we described above.

Our results show that it is possible to identify a privileged speaker from a small group in near real time and to apply speaker dependent speech processing technologies to the resulting utterances.



**Fig. 6.** The primary-speaker indicator shows a positive identification from raw speech derived in real time. This indicator is generated on a regular basis and allows a speech time series to be segmented into portions generated by the privileged speaker and other portions.



**Fig. 7.** Real time non-privileged speaker indicator. This indicator allows speech or noise segments that not associated with the privileged speaker to be identified.



## 5 Conclusions

We have shown a proof-of-concept for a real time *user sensitive interface* that could, in the not too distant future, allow each individual in a working group to be served by an attentive interface that can select the speech from its primary, or privileged, speaker and apply speaker dependent speech recognition to his or her utterances. These specific utterances can be parsed for commands that would be applied to a context sensitive interface for that person, or to collect dictated meeting minutes. The NIST Mk-III microphone array can be used, with phased array processing, to find the bearing of a speaker using a *steered response beamformer* and then use the bearing to acquire speech at good signal-to-noise levels for further spoken language processing. Moreover, the speaker identification algorithm is effective over a range of bearing angles. This, or functionally similar technology will undoubtedly form a foundational element of context-aware spoken interface systems that can serve the needs of diverse individuals, even those with special needs.

We must caveat this by noting that this simple proof-of-concept needs improvements in many areas to make it fully practical. First, we must improve signal-to-noise ratio of the acquired speech signals by implementing an adaptive beamforming algorithm to acquire speech from physically plausible tracks. Because people in conference rooms move slowly, at least in terms of microphone array frames at 44 kHz, we can constrain the search space to places near the previous sources. As a result, the speaker verification performance will be enhanced. Second, the speech activity detection and segmentation is currently carried out by a simple instantaneous threshold selection procedure. It must be investigated how to segment speech signals based on utterance in noisy environments. Another issue that must be improved is that of source separation. NIST Meeting Room Recognition project studies show that nearly seventy percent of spontaneous speech segments contain overlapped speech by multiple speakers [18, 19]. This implies a need for research in source separation in a sound field measured by many microphones simultaneously. Some combination of beamforming and source separation will probably be a key technology element for robust meeting speech recognition for the future.

Intelligent Automation wishes to acknowledge support by the National Institute of Standards Information Technology Laboratory, and its Technology Small Business Innovative Research Program, under contract number: SB1341-02-W-1140 for this work.

## References

- [1] Stanford V., Smart Space Scenario. Proceedings of the 1998 DARPA/NIST Smart Spaces Workshop, July 30-31, Gaithersburg, MD (1998) 1.1-1.2
- [2] Flanagan J. and Stanford V., Situation Awareness in Smart Spaces. Proceedings of the 1998 DARPA/NIST Smart Spaces Workshop, July 30-31, 1998, Gaithersburg, MD (1998) 3.1-3.13
- [3] Li Q. and Juang B., Speaker Authentication. Pattern Recognition in Speech and Language Processing, W. Chou, B. Juang, (eds.) CRC Press (2003) 229-259

- [4] Reynolds D. and Rose R., Robust Text-Independent Speaker Verification Using Gaussian Mixture Speaker Models. *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, (1995)
- [5] Kwan C. et al., A Real-Time Demonstration of the NIST Smart Flow System, Phase 1 SBIR Final Report, (2003)
- [6] Flanagan J., Berkley D., Elko G., West J., and Sondhi M., Autodirective Microphone Systems. *Acustica*, pages, Vol. 73, (1991) 58-71
- [7] DeGraaf S. and Johnson D., Capability of Processing Algorithms to Estimate Source Bearings. *IEEE Trans. On Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 6 (1985) 1368-1379
- [8] Johnson D. and DeGraaf S., Improving the Resolution of Bearing in Passive Sonar Arrays by Eigenvalue Analysis. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 6, (1982) 638-647
- [9] Omologo M., Matassoni M., and Svaizer P., Speech Recognition with Microphone Arrays. *Microphone Arrays*. In: Brandstein M. and Ward D. (eds.) *Signal Processing Techniques and Applications*. Springer-Verlag, Berlin, Heidelberg, New York (2001) 331-349
- [10] Flanagan J. and Huang T. (eds.), Special Issue on Human-Computer Multimodal Interface, *Proc. of the IEEE*, Vol. 91, No. 9, Sept. 2003
- [11] Hazen T., et al. A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments. *Proc. of the Sixth International Conference on Multimodal Interfaces*, October 14-15, State College, Pennsylvania, USA (2004) 235-242
- [12] Rose R., Quek F., and Shi Y., MacVisSTA: A System for Multimodal Analysis. *Proc. of the Sixth International Conference on Multimodal Interfaces*, October 14-15, State College, Pennsylvania, USA (2004) 259-264
- [13] Demirdjian D., Wilson K., Siracusa M., and Derrell T., Real-time Audio-Visual Tracking for Meeting Analysis. *Proc. of the Sixth International Conference on Multimodal Interfaces*, October 14-15, State College, Pennsylvania, USA (2004) 331-332
- [14] Rabiner L. and Juang B-H., Linear Predictive Coding Model for Speech Recognition. In: *Fundamentals of Speech Recognition*, PTR Prentice-Hall Inc. Englewood Cliffs, New Jersey, USA (1993) 97-121
- [15] Knill K. and Young S., Hidden Markov Models in Speech and Language Processing. In: *Corpus-Based Methods in Language and Speech Processing*. Young S. and Bloothoft G., (eds.), Kluwer Academic Publishers, Norwell, MA, USA pages (1997) 36-41
- [16] Parzen E., On estimation of a probability density function and mode. *Ann. Math. Stat.* Vol. 33, (1962) 1065-1076
- [17] Nabney I., *Netlab Algorithms for Pattern Recognition*. Springer, New York, (2001)
- [18] Fiscus J., Radde N., Garofolo J., Le A., Ajot J., Laprun C., 2005 The Rich Transcription Spring Meeting Recognition Evaluation: NIST MLMI Meeting Recognition Workshop, Renals S. and Bengio S., (eds). Edinburgh. To appear in *Springer Lecture Notes in Computer Science Series*, Volume 3869
- [19] Fiscus J., Ajot J., Radde N., Laprun C., Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. To appear: LREC, May 2006, Genoa, Italy

# A Middleware for the Deployment of Ambient Intelligent Spaces

Diego López-de-Ipiña, Juan Ignacio Vázquez, Daniel Garcia, Javier Fernández,  
Iván García, David Sáinz and Aitor Almeida

University of Deusto, Faculty of Engineering, Avda. Universidades 28,  
48007 Bilbao, Spain

{dipina, ivazquez}@eside.deusto.es,  
{dsainz, aalmeida}@tecnologico.deusto.es

**Abstract.** The latest mobile devices are offering more multimedia features, better communication capabilities (Bluetooth, Wi-Fi, GPRS/UMTS) and are more easily programmable than ever before. So far, those devices have been used mainly for communication, entertainment, and as electronic assistants. On the other hand, Ambient Intelligence (AmI) is emerging as a new research discipline merging the fields of Ubiquitous Computing and Communications, Context Awareness and Intelligent User Interfaces. The ultimate goal of AmI is to surround our working and living environments with context-aware, cooperative and invisible devices that will assist and help us in our everyday activities. Current mobile devices, which accompany us anywhere and at anytime, are the most convenient tools to help us benefit from AmI-enhanced environments. In other words, mobile devices are the best candidates to intermediate between us and our surroundings. In consequence, this paper proposes a middleware which aims to make this vision reality following a two-fold objective: (1) to simplify the creation and deployment of physical spaces hosting smart objects and (2) to transform mobile devices into universal remote controllers of those objects.

## 1 Introduction

Ambient Intelligence (AmI) [13] defines an interaction model between us and a context-aware environment, which adapts its behaviour intelligently to our preferences and habits, so that our life is facilitated and enhanced.

Current PDAs and mobile phones are equipped with significant processing and storage capabilities, varied communications mechanisms (Bluetooth [1], Wi-Fi, GPRS/UMTS) and increasingly capable multimedia capture and playback facilities. Moreover, they are far more easily programmable (Compact.NET [9], J2ME [15] or Symbian [17]), i.e. extensible, than ever before.

Mobile devices equipped with Bluetooth, built-in cameras, GPS receivers, barcode or RFID readers can be considered as sentient devices [8][11], since they are aware of what smart objects are within an AmI space. By Smart Space (or AmI-enhanced environment), we understand a location, either indoors or outdoors, where the objects present within (smart objects) are augmented with computing services. A smart object is an everyday object (door, classroom, parking booth) or a device augmented with some accessible computational service [2]. Once a mobile device discovers a nearby smart object, it may operate over it.

We deem that mobile devices will play a key role within AmI, since they are always with us and can act as facilitators or intermediaries between us and the environment. In other words, mobile devices can act as our personal electronic butlers, facilitating and enhancing our daily activities, and even acting on our behalf based on our profiles or preferences.

In this paper, we describe the design and implementation of EMI<sup>2</sup>lets, a middleware to facilitate the development and deployment of mobile context-aware applications for AmI spaces. This software provides the software infrastructure to (1) convert physical environments into AmI spaces and (2) transform mobile devices into remote controllers of the smart objects in those spaces.

The structure of the paper is as follows. Section 2 describes EMI<sup>2</sup>, a software architecture modelling both passive and active interaction mechanisms for AmI. Section 3 introduces the EMI<sup>2</sup>lets platform, a partial materialisation of the EMI<sup>2</sup> architecture, which simplifies both the creation of software representatives for everyday objects and their controlling proxies deployable in mobile devices. Section 4 illustrates the life cycle of an EMI<sup>2</sup>let from its development to its deployment in a mobile device and lists some interesting applications produced with the EMI<sup>2</sup>lets platform. Section 5 shows some performance results achieved by the current implementation of EMI<sup>2</sup>lets. Section 6 overviews some related work. Finally, section 7 offers some conclusions and suggests further work.

## 2 EMI<sup>2</sup>: An AmI Architecture

Regardless of the continuous progress achieved in all the related research topics that contribute to the AmI vision, we are still far away from its materialisation. A good starting point to solve this may be the definition of suitable software architectures and frameworks specially catered for AmI. The EMI<sup>2</sup> (Environment to Mobile Intelligent Interaction) architecture is our proposed solution.

EMI<sup>2</sup> defines a multi-agent software architecture, where agents of different types, modelling the different roles played by entities in AmI, communicate and cooperate to fulfil a common goal, i.e. to enhance and facilitate the user interactions with her *AmI Space*. For instance, a cinema may be enhanced with a Bluetooth mobile phone accessible ticket booking service, preventing the user from long queuing to purchase tickets. Similarly, the door of our office may be augmented with an access control service, which demands the user to enter a PIN in her mobile to be given access.

Fig. 1 portrays the main components of the EMI<sup>2</sup> architecture. We distinguish three main types of agents:

- *EMI<sup>2</sup>Proxy*: is an agent representing the user, which runs on the user's mobile device (PDA or mobile phone). It acts on behalf of the user, adapting/controlling the environment for him, both *explicitly*, under the user's control, or *implicitly*, on its own judgement based on profiles, preferences and previous interactions.
- *EMI<sup>2</sup>Object*: is an agent representing any device or physical object (vending machine, door, ticket box) within a smart environment augmented with computational services, i.e. the capacity to adapt its behaviour based on ambient conditions or user commands. An EMI<sup>2</sup>Object cooperates with other EMI<sup>2</sup> agents.

- *EMI<sup>2</sup>BehaviourRepository*: is an agent where knowledge and intelligence are combined to support sensible adaptation. EMI<sup>2</sup>Objects may require the assistance of an external EMI<sup>2</sup>BehaviourRepository to coordinate their own adaptation according to the user's preferences, behaviour patterns or even the explicit commands received from an *EMI<sup>2</sup>Proxy*. The user's mobile device can also be powered with an internal EMI<sup>2</sup>BehaviourRepository.

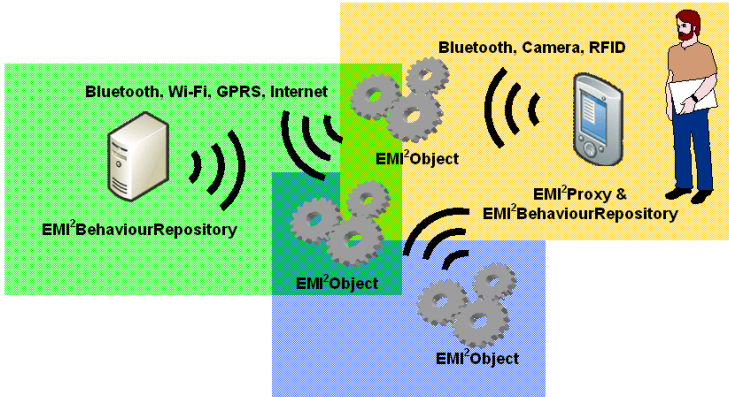


Fig. 1. The EMI<sup>2</sup> Architecture

## 2.1 Active and Passive Mechanisms

A concrete agent can influence the environment, and thus, its constituent agents' state, via *active* (explicit interaction) or *passive* (implicit interaction) methods.

Active methods are those in which the agent explicitly commands other agents to change their state or perform an action. For example, when a user enters a building, a sensor identifies him and commands the lift to be ready at the ground floor. When the user stands by his office door, his mobile phone commands the electric lock to open. Active methods can be implemented with any distributed computing technology capable of issuing commands, which will be transported in a local context by bearers such as Bluetooth or Wi-Fi and in a global context by GPRS/UMTS.

Passive methods [19] are those in which an agent disseminates certain information (profiles, preferences), expecting that other agents change their state or perform an action at their discretion to create a more adapted environment. Using passive methods, an agent does not command the target agents to do anything concrete. It simply publishes/broadcasts information preferences, expecting the others to react by changing their state in a positive way. Passive mechanisms are less intrusive than active methods, but they are less predictable and significantly more complex.

## 2.2 Active Influence over EMI<sup>2</sup>Objects

In this paper, we want to concentrate on the design and implementation of a middleware to provide universal active influence capabilities to our mobile devices over the surrounding smart objects in our environment.

The minimum features such a middleware has to provide are: (1) a mechanism to discover, through ad-hoc or wireless networking, the computing services exported by surrounding smart objects, and (2) a mechanism to interact with those discovered services, so that the objects they represent adapt to the user's preferences and commands.

The current state of the art in discovery and interaction platforms falls into three categories [5][21]. Firstly, solutions in which discovery protocols are supported by mobile code, e.g. Jini [16]. After discovery, the service (either a proxy or the full service) is downloaded onto the mobile device where it then operates. Secondly, solutions where the discovery protocols are integrated with specific interaction protocols, which are used to invoke the service after the service has been discovered. A good example of this is Universal Plug and Play (UPnP) [18]. Finally, there are interaction independent discovery protocols such as the Service Location Protocol [3].

In what follows, we explain the design and implementation of an AmI-enabling middleware that addresses the service discovery and interaction aspects required for active influence (explicit invocation) on  $EMI^2$ Objects.

### 3 The $EMI^2$ lets Platform

$EMI^2$ lets is the result of mapping the  $EMI^2$  architecture into a software development platform to enable AmI scenarios. This platform is specially suited for active interaction mechanisms. However, it has been designed so that passive mechanisms may be incorporated in the future.

$EMI^2$ lets is a development platform for AmI which addresses the intelligent discovery and interaction among  $EMI^2$ Objects and  $EMI^2$ Proxies.  $EMI^2$ lets follows a Jini-like mechanism by which once a service is discovered, a proxy of it (an  $EMI^2$ let) is downloaded into the user's device ( $EMI^2$ Proxy). An  $EMI^2$ let is a mobile component transferred from a smart object to a nearby handheld device, which offers a graphical interface for the user to interact over that smart object.

The  $EMI^2$ lets platform addresses three main aspects:

- *Mobility*, seamlessly to the user, it encounters all the services available as he moves and selects the best possible mechanism to communicate with them. In other words, the  $EMI^2$ let platform ensures that an  $EMI^2$ Proxy is always using the communication means with best trade-off between performance and cost. For example, if Wi-Fi and Bluetooth are available, the former is chosen.
- *Interoperability*, the  $EMI^2$ lets are agnostic of the target device type, e.g. PC, a PDA or a mobile phone.
- *AmI* is the application domain that has driven the design of  $EMI^2$ lets. This platform provides the infrastructure and software tools required to ease the development and deployment of AmI scenarios.

The objectives established for the design and implementation of the  $EMI^2$ lets platform are:

- Transform mobile devices (mobile phones and PDAs) into remote universal controllers of the smart objects located within an AmI space.

- Enable both local (Bluetooth, Wi-Fi) and global access (GPRS/UMTS) to the smart objects in an AmI space, seamlessly adapting to the most suitable underlying communication mechanisms
- Develop middleware independent of a particular discovery or interaction mechanism. Abstract the programmer from the several available discovery (Bluetooth SDP or wireless UPnP discovery) and interaction mechanisms (RPC or publish/subscribe). Allow this middleware to easily adapt to newly emerging discovery (e.g. RFID identification) and interactions means.
- Make use of commonly available hardware and software features in mobile devices, without demanding the creation of proprietary hardware, or software.
- Generate software representatives (proxies) of smart objects which can be run in any platform, following a “write once run in any device type” philosophy. For instance, the same EMI<sup>2</sup>let could be run in a mobile, a PDA or a PC.

### 3.1 The EMI<sup>2</sup>lets Vision

Fig. 2 shows a possible deployment of an EMI<sup>2</sup>let-aware environment. A group of devices running the EMI<sup>2</sup>let Player and hosting the EMI<sup>2</sup>let runtime can discover and interact with the software representatives (EMI<sup>2</sup>lets) of surrounding EMI<sup>2</sup>Objects. An EMI<sup>2</sup>Object may be equipped with enough hardware resources to host an EMI<sup>2</sup>let Server, or alternatively, a group of EMI<sup>2</sup>lets associated to different EMI<sup>2</sup>Objects may all be hosted within an autonomous version of an EMI<sup>2</sup>let Server. The EMI<sup>2</sup>let Server acts as a repository of EMI<sup>2</sup>Objects. It publishes the services offered by the hosted EMI<sup>2</sup>Objects, transfers them on demand to the requesting EMI<sup>2</sup>let Players, and, optionally acts as a running environment for the EMI<sup>2</sup>let server-side facets.

Some EMI<sup>2</sup>lets may directly communicate with their associated EMI<sup>2</sup>Objects in order to issue adaptation commands. However, often a specialised piece of software may need to be developed which is far too complex to be implemented in the embedded hardware with which a smart object is normally equipped. For those cases, it will be more convenient to delegate those cumbersome computing tasks to the server-side (back-end) counterpart of an EMI<sup>2</sup>let. The EMI<sup>2</sup>let on the hand-held device will communicate with its server-side counterpart in the EMI<sup>2</sup>let Server by means of the EMI<sup>2</sup>Protocol. For example, a light-controlling EMI<sup>2</sup>let could communicate with its EMI<sup>2</sup>let server-side, which would issue X10 commands over the power line.

### 3.2 Internal Architecture

The EMI<sup>2</sup>lets platform consists of the following elements:

1. A programming framework defining a set of classes and rules that every EMI<sup>2</sup>let component must follow.
2. An integrated development environment, named EMI<sup>2</sup>let Designer, which simplifies the development of EMI<sup>2</sup>lets, both its client- and (optional) server-side.
3. A runtime environment installed on EMI<sup>2</sup>let-aware devices for executing downloaded code.

4. An EMI<sup>2</sup>let Player to discover, download, verify and control the execution life of a downloaded EMI<sup>2</sup>let. A version of the player is available for each device type which may act as host of EMI<sup>2</sup>lets, e.g. PDA, mobile phone or PC.
5. An EMI<sup>2</sup>let Server which acts as repository of EMI<sup>2</sup>lets and as running environment of EMI<sup>2</sup>lets server-sides.

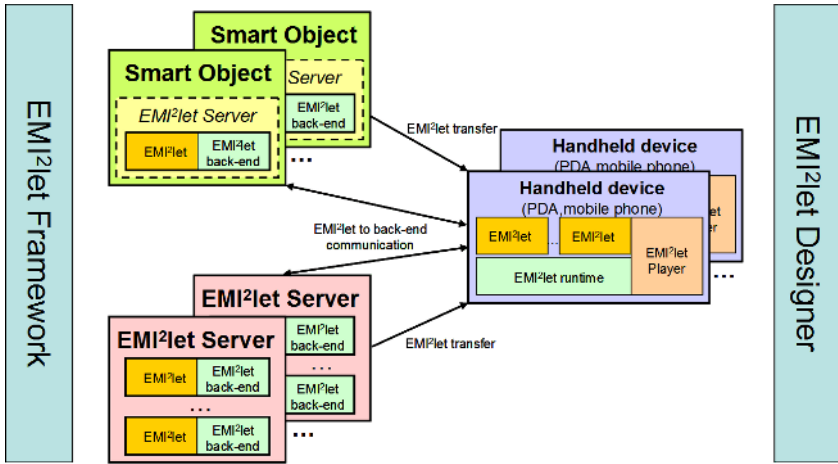


Fig. 2. The EMI<sup>2</sup>lets in action

In order to achieve the design objectives previously listed, we have created the layered software architecture shown in Fig. 3. Programmers only deal with the first layer, the *EMI<sup>2</sup>let Abstract Programming Model API*, to develop the software counterparts of smart objects. This layer offers a set of generic interfaces (abstract classes) covering the main functional blocks of a mobile sentient application:

1. *Discovery* interface to undertake the search for available EMI<sup>2</sup>lets independently of the discovery mechanisms used underneath.
2. *Interaction* interface to issue commands over the services discovered.
3. *Presentation* interface to specify the graphical controls and events that represent the look and feel of an EMI<sup>2</sup>let.
4. *Persistency* interface to store EMI<sup>2</sup>let-related data in the target device.

The *EMI<sup>2</sup>let Abstract-to-Concrete Mapping* layer translates the invocations over the generic interfaces to the appropriate available mechanisms both in the mobile device and the EMI<sup>2</sup>Objects in the environment. Abstractions encapsulate the concrete discovery, interaction, presentation or persistency models use discovery, interaction, presentation and persistency. They implement an API for performing service discovery and interaction, graphical interface generation and data persistence independent of the actual implementation in the target device. On deployment, the code generated through these abstract interfaces is linked to the concrete implementations of the abstractions which are part of the EMI<sup>2</sup>let runtime in the target device.



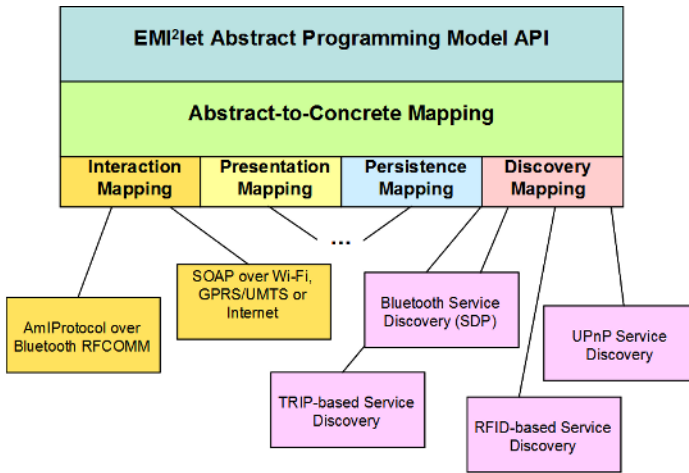


Fig. 3. EMI²lets Internal Architecture

In the process of associating a generic invocation to an actual one, the *EMI²let Abstract-to-Concrete Mapping* will be responsible of selecting the actual mapping (or group of mappings) which best matches the invocation type. For example, if a downloaded EMI²let is installed on a device where both Bluetooth and GPRS communication are available, the abstract-to-concrete layer will have to choose one of those mechanisms to issue commands. Thus, if the mobile device is still within Bluetooth range of the EMI²let server-side, then it will translate the invocation into an EMI²Protocol message transported over Bluetooth RFCOMM. Otherwise, it will invoke via GPRS the generic web service (with methods corresponding to the EMI²Protocol commands) implemented by every EMI²let server-side.

Similarly, if a mobile device is Bluetooth and Wi-Fi capable, it will use both Bluetooth SDP and UPnP service discovery to concurrently search for smart objects in its surroundings.

With regards to the presentation abstraction, we have defined a minimum set of graphical controls with which we can generate the graphical interface of an EMI²let. Currently, we support the following controls: panel, label, button, textbox, checkbox, combobox, listbox, sound and image. Some examples of the graphical control classes defined are: *EMI²Panel*, *EMI²Button* or *EMI²TextBox*. This set of controls enables us to create graphical interfaces for EMI²lets which are agnostic to the target mobile device. Thus, when a programmer creates an *EMI²Button*, it is not translated into a button control in a PC or a PDA, but into a menu option in a mobile phone. Still, in order to guarantee a proper layout of the graphical controls according to the three target device types (PC, PDA and mobile phone) supported, specific layout hints can be given, with the help of the EMI²let Designer, for each device type. An example showing this fact can be seen in Fig. 4. Lately, we have added support for a new target device type, namely web-enabled devices. We have enhanced the functionality of the EMI²let Server so that it can export EMI²lets as web pages accessible from non EMI²let-compliant web-enabled devices.

The modus operandi of the plug-ins associated to any of the four available functional mapping is ruled by an XML configuration file, which states whether a plug-in may be run concurrently with other plug-ins of the same type or in isolation. In the latter case, a priority is assigned to each plug-in which will determine which of the plug-ins to select when several of them are available. We plan to establish a more sophisticated and flexible plug-in configuration model in due time.

Both the *Abstract-to-Concrete Mappings* and the *Functional Mapping* (plug-ins) layers compose the runtime installed in each target device. The code of the downloaded EMI<sup>2</sup>let is linked on arrival by the EMI<sup>2</sup>let Player with the runtime.

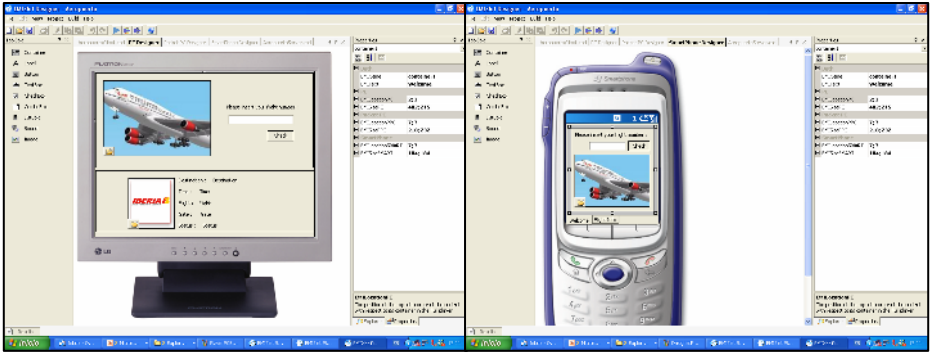


Fig. 4. EMI<sup>2</sup>lets Designer

### 3.3 The EMI<sup>2</sup>Protocol

The EMI<sup>2</sup>Protocol defines a set of basic commands or messages which are exchanged among EMI<sup>2</sup>let Players and Servers and EMI<sup>2</sup>let client- and server-sides. For those messages to be exchanged, a connection between the client and server peers must have previously been established. The discovery plug-ins in a Player are in charge of discovering surrounding Servers and opening connections between Players and Servers.

The plug-in implementations which use the Bluetooth bearer actually exchange the commands as specified below, whereas plug-ins using other bearers such as Wi-Fi or GPRS invoke a generic web service with methods corresponding to those commands. The most important commands offered are:

- HELLO, through this message an EMI<sup>2</sup>let Player provides metadata to an EMI<sup>2</sup>let Server. The metadata provided is the type of device the Player is running on and the set of communication and discovery mechanisms (i.e. plug-ins installed and running) available at the device.
- HELLO\_RESPONSE, the EMI<sup>2</sup>let Server informs the connected EMI<sup>2</sup>let Player about the communication channels it supports (i.e. the server's own plug-ins installed and running).
- SERVICE\_QUERY, message issued by an EMI<sup>2</sup>let Player to an EMI<sup>2</sup>let Server in order to retrieve information about the EMI<sup>2</sup>lets it hosts.

- `SERVICE_QUERY_RESPONSE`, message issued by an EMI<sup>2</sup>let Server to provide a requesting EMI<sup>2</sup>let Player with the metadata of the services it hosts.
- `DOWNLOAD_SERVICE`, message issued by an EMI<sup>2</sup>let Player to an EMI<sup>2</sup>let Server in order to retrieve the code of a selected EMI<sup>2</sup>let.
- `DOWNLOAD_SERVICE_RESPONSE`, message issued by an EMI<sup>2</sup>let Server to provide a requesting EMI<sup>2</sup>let Player with the code of a selected service.
- `COMMAND_SEND`, message encapsulating a packet of data sent between an EMI<sup>2</sup>let executing in a Player and its server-side hosted on an EMI<sup>2</sup>let Server.
- `COMMAND_RESPONSE`, message encapsulating a packet of data sent between an EMI<sup>2</sup>let server-side and the EMI<sup>2</sup>let running in the Player.

### 3.4 Implementation

Reflection is paramount in the EMI<sup>2</sup>lets platform. It enables an EMI<sup>2</sup>let Player to verify that the code arriving as part of an EMI<sup>2</sup>let complies with the EMI<sup>2</sup>lets framework and can be trusted. Every EMI<sup>2</sup>let downloaded is encrypted with a private key only shared by the EMI<sup>2</sup>let designer and the player. After downloading an EMI<sup>2</sup>let, the Player unencrypts its code and verifies that the class downloaded follows the EMI<sup>2</sup>let framework rules.

After verification, the player can start the EMI<sup>2</sup>let by invoking the methods defined in the `EMI2let` base class, from which every EMI<sup>2</sup>let must inherit. The methods defined by this class closely resemble the ones provided by a J2ME [9] `MIDlet` class:

- `Start`, starts or resumes the execution of a downloaded EMI<sup>2</sup>let.
- `Pause`, pauses its execution.
- `Destroy`, destroys it.

In addition, the `EMI2let` class includes some EMI<sup>2</sup>let-specific methods such as:

- `GetUUID`, returns the unique identifier of an EMI<sup>2</sup>let.
- `SetProperty/GetProperty`, sets or gets the properties associated to a EMI<sup>2</sup>let. For instance, the `EMI2let.Durable` property is set to `true` when an EMI<sup>2</sup>let has to be cached in the player after its execution. Thus, it can be executed again in the future. Otherwise, an EMI<sup>2</sup>let is wiped out from the Player either when its execution is completed or it is out of range of the EMI<sup>2</sup>Object it represents.
- `NotifyDisconnected`, offers an EMI<sup>2</sup>let the possibility of being aware of when the EMI<sup>2</sup>Object that it controls cannot be accessed any longer.
- `GetAddresses`, enables the EMI<sup>2</sup>let-hosting player to retrieve the addresses at which the EMI<sup>2</sup>let server-side is available. For instance, an EMI<sup>2</sup>let server-side may be accessible both through a Bluetooth address or a url pointing to a web service.

Our first reference implementation has used Microsoft .NET, a platform that fully supports reflection through the `System.Reflection` namespace. Moreover, the .NET platform addresses software development for all the client hardware platforms considered in EMI<sup>2</sup>lets, namely PC, PDA and mobile phone. As a least common multiple for the definition of the presentation controls of an EMI<sup>2</sup>let, we have chosen most of the Compact.NET framework graphical controls, which represent a superset

of the ones in the SmartPhone framework and a subset of the standard .NET desktop-oriented ones.

The most noticeable part of our implementation is the assembly fusion undertaken at the player side merging the arriving EMI<sup>2</sup>let assembly with the EMI<sup>2</sup>let library installed in each target device. This library represents the player's runtime, i.e. the abstract-to-concrete layer and the interaction, discovery, presentation and persistency mappings implementation with their corresponding plug-in modules. In other words, the assembly code downloaded is linked dynamically (late bound) with the runtime installed in the target device. This assembly process would not have been possible without the use of reflection.

### 3.5 An EMI<sup>2</sup>let Plug-In Example

The plug-in based mechanism adopted in EMI<sup>2</sup>lets guarantees its extensibility. If we want to add support to EMI<sup>2</sup>lets for any newly emerging discovery or communication technology, we simply need to implement a plug-in for the corresponding abstraction. In order to prove this point, we have implemented a service discovery plug-in based on the TRIP [7] tag-based visual identification system.

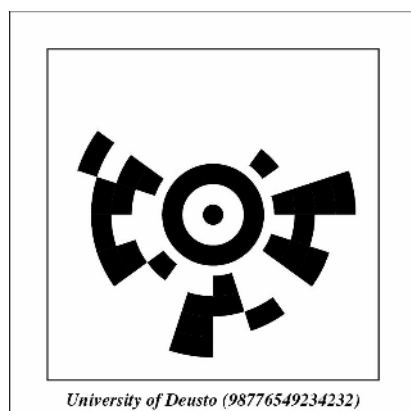
A factor that limits the use of Bluetooth as an underlying networking technology for publicly accessible mobile services is that its device discovery process takes a significant (sometimes unbearable) time. The discovery process in Bluetooth is divided into two main phases: (1) device discovery, i.e. what other devices are accessible via Bluetooth, and (2) service discovery, i.e. what services are offered by the discovered devices. In an error-free environment, the device discovery phase must last for 10.24s if it is to discover all the devices [1].

In order to speed up service discovery, we have devised a tag-based content/service selection mechanism, which bypasses the slow Bluetooth device discovery process. Our approach is inspired by the work of [12].

The TRIP visual tags are circular barcodes (*ringcodes*) with 4 data-rings and 20 sectors. A visual tag, large enough to be detected by a mobile device tag reading software, is shown in Fig. 5. The ringcode is divided into: (1) one *sync-sector* used to specify the beginning of the data encoded in a tag, (2) two *checksum-sectors* used to encode an 8-bit checksum, which detects decoding errors and corrects three bit errors, and (3) seventeen *data-sectors*, which encode 66 bits of information.

The information in a TRIP tag is encoded in anti-clockwise fashion from the sync sector. Each sector encodes a hexadecimal digit comprising the values 0 to D. The E hexadecimal number is only permitted in the sync sector. Given the 17 data encoding sectors, the range of valid IDs is from 0 to  $15^{17}-1$  ( $98526125335693359375 \approx 2^{66}$ ).

The TRIP tags were designed to work well with the low-resolution fixed-focal-length cameras found on conventional CCTV systems. Consequently, they are also very well suited for the low-quality built-in cameras of mobile devices, as we suggested in [8]. In fact, our experience shows that the TRIP ringcodes are more reliably recognized than linear (UPC) barcodes, which demand far higher image resolutions. TRIP works reliably with 160x120 pixel images taken at a distance of 5-30 cms from the tags that label the smart objects in an environment. We have implemented the TRIP tag reading software for Compact.NET devices, achieving 2 fps in a TSM 500 Pocket PC.



**Fig. 5.** A tag encoding 66 bits of data

### 3.6 EMI<sup>2</sup>lets State Management

The EMI<sup>2</sup>lets platform incorporates a simple state management mechanism. In order to prevent the user from continuously entering the same input details in the execution of a previously run (durable) EMI<sup>2</sup>let, the player stores in EMI<sup>2</sup> Cookie objects the last values input. The EMI<sup>2</sup> Cookies, contrary to the well-known HTTP cookies, keep the state in the player (client-side). The UDDI associated to an EMI<sup>2</sup>let is employed to establish associations between EMI<sup>2</sup>lets and EMI<sup>2</sup>Cookies. Currently, we are working on extending state management in EMI<sup>2</sup>lets by adopting the WebProfiles model proposed at [20].

## 4 EMI<sup>2</sup>lets Applications

In this section, we will first describe the lifecycle of an EMI<sup>2</sup>let from its development to its deployment and secondly, we will mention some of the applications developed with EMI<sup>2</sup>lets.

### 4.1 The Life Cycle of an EMI<sup>2</sup>let

Fig. 6 shows the life cycle of an EMI<sup>2</sup>let from its development with the EMI<sup>2</sup>let Designer (see Fig. 4) until its deployment at the target mobile device and EMI<sup>2</sup>let server. In our approach, active .NET code developed on a PC with the help of the EMI<sup>2</sup>let Designer is uploaded into an EMI<sup>2</sup>let Server, from where it is later discovered, downloaded, verified and executed in the context of an EMI<sup>2</sup>let Player. After its execution and depending on its durability properties, the EMI<sup>2</sup>let is cached or removed from the Player.

### 4.2 Examples of EMI<sup>2</sup>lets

We have developed EMI<sup>2</sup>lets targeted to the following application domains: accessibility, home/office automation, industry and public spaces.

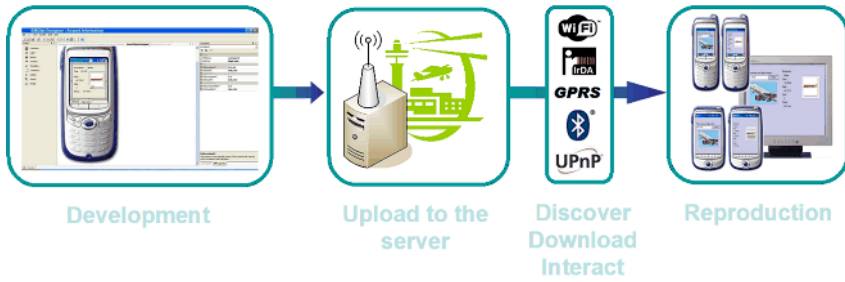


Fig. 6. EMI<sup>2</sup>let Life Cycle

In the domain of accessibility we have developed EMI<sup>2</sup>lets which, associated to a bus stop, offer a voice synthesized bus arrival notification for blind people or provide subtitles on the mobile phones of people attending to a conference. These applications demonstrated how simple it is to transform a physical space (bus stop or conference hall) into a more accessible environment thanks to the EMI<sup>2</sup>lets platform.

In the home and office automation domain we have created EMI<sup>2</sup>lets that enable us to control from our mobile devices the lights, music system (in fact the Windows Media Player in a PC) and a Pan/Tilt/Zoom security camera at a home or office.

As far as the industry domain is concerned, we have developed an EMI<sup>2</sup>let which allows us to control from our mobile device a robot equipped with a communications module supporting both Bluetooth and GPRS. When co-located with the robot, our EMI<sup>2</sup>let uses the Bluetooth communication channel. When we are far away from the location of the robot, the EMI<sup>2</sup>let uses the GPRS channel to communicate with the robot. The communication channel choice is undertaken by the EMI<sup>2</sup>lets runtime autonomously.

Finally, on what we call the “public space” domain, we have created EMI<sup>2</sup>lets which allow us to control a parking booth, order food in a restaurant or review the departure time and gate of a plane in an airport. Those EMI<sup>2</sup>lets show how a physical object in an outdoors space can be augmented with AmI features. For example, the Parking EMI<sup>2</sup>let is meant to be deployed in any street parking booth, where we can purchase tickets to park our car for a limited period of time. Often, we have to keep returning to the parking place to renew the ticket so that the local police force does not issue a fine for parking time expiration. Thanks to the EMI<sup>2</sup>lets platform, a user could discover, download (from the ticket booth) and install a parking EMI<sup>2</sup>let which would help him solve this situation. With the downloaded EMI<sup>2</sup>let, the user could purchase parking tickets via Bluetooth while in the parking, and remotely via GPRS when the EMI<sup>2</sup>let warns her (at her office) that its parking ticket is about to expire. This scenario shows one of the biggest virtues of EMI<sup>2</sup>lets, namely its capability to enact an action over an EMI<sup>2</sup>Object both locally, while in the environment, or remotely, far away from the environment.

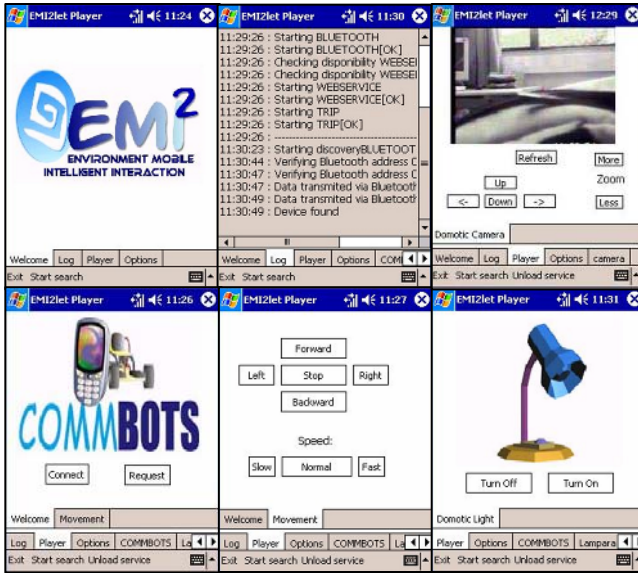


Fig. 7. EMI<sup>2</sup>lets running on a PDA

Fig. 7 and Fig. 8 show three of the previously described EMI<sup>2</sup>lets in action running in a PDA and a mobile phone, respectively. The EMI<sup>2</sup>lets shown allow a user to control from his mobile device a robot, a lamp or a PTZ security camera. Something remarkable about the EMI<sup>2</sup>lets platform is that in the development of those EMI<sup>2</sup>lets we have written the code only once, independently of the target device where they will run. This is due to the “write once run in any device type” philosophy followed by our system.

## 5 EMI<sup>2</sup>lets Performance Results

In order to assess the performance of our current implementation of the EMI<sup>2</sup>lets platform, we have carried out two tests:

1. A comparative measurement illustrating the different latencies experienced during an EMI<sup>2</sup>let discovery, download and communication with its server-side, bearing in mind the nature of the communication channel used (Wi-Fi, Bluetooth or GPRS).
2. A comparative measurement to determine the average data rate achieved depending on whether we use Bluetooth, Wi-Fi or GPRS to transfer data between an EMI<sup>2</sup>let and its server-side.

Fig. 9 shows that the discovery process based on UPnP over Wi-Fi is much faster than connecting directly to the IP address and port number of an EMI<sup>2</sup>let Server to enquire about its installed EMI<sup>2</sup>lets over GPRS or undertaking Bluetooth discovery. However, once the Bluetooth discovery has concluded the download of an EMI<sup>2</sup>let code and the exchange of information between an EMI<sup>2</sup>let and its server-side is much better than through GPRS and only worse to Wi-Fi which has a much better transfer rate.

Fig. 10 shows the effective data transfer rates obtained over the three wireless communication mechanisms we have used in EMI<sup>2</sup>lets. Obviously, the data transfer rate obtained through Wi-Fi is the best, whereas Bluetooth offers the second best behaviour.



Fig. 8. EMI<sup>2</sup>lets running on a mobile phone

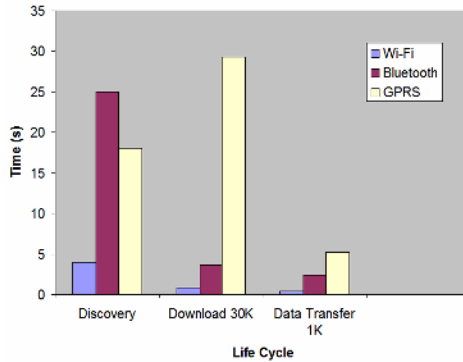


Fig. 9. EMI<sup>2</sup>lets communication costs

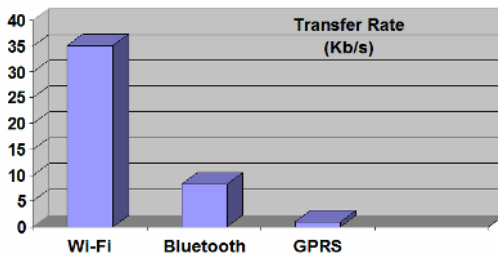


Fig. 10. Effective data transfer rate in EMI<sup>2</sup>lets



## 6 Related Work

The EMI<sup>2</sup>lets platform presents some resemblance to the Smoblets software framework proposed by [14]. Both frameworks download into a mobile device and the software representatives of objects are located in a smart space. However, Smoblets only operate when they are within range of the smart object they represent. On the contrary, EMI<sup>2</sup>lets can remain at the user's terminal, even when he is far away from the smart object. This allows the user to control that smart object anytime and anywhere, both using local (Bluetooth) and global (GPRS) communication mechanisms. Furthermore, the main application of Smoblets is to transform mobile devices into execution platforms for code downloaded from smart items with limited processing resources, whereas EMI<sup>2</sup>lets are mainly thought to transform mobile devices into hosts of smart object proxies, which simplify their remote control.

The EMI<sup>2</sup>lets framework's layered software architecture has been inspired by the ReMMoC framework [5]. However, EMI<sup>2</sup>lets does not only address the service discovery and interaction issues of mobile context-aware applications. It also tackles the graphical presentation and persistency aspects commonly used in those applications. Moreover, the EMI<sup>2</sup>let code generated is independent of the target platform where it will be run (PC, PDA or mobile).

The Pebbles project [10] is exploring how handheld devices, such as PDAs and mobile phones, can be used when they are communicating with a "regular" personal computer (PC), with other handhelds, and with computerized appliances such as telephones, radios, microwave ovens, automobiles, and factory equipment. Pebbles shares with EMI<sup>2</sup>lets the goal of transforming handheld devices into universal remote controllers. Moreover, it adopts a similar architecture where a player in the handheld device communicates with server-side intermediaries to control the operation of the underlying smart objects. However, the main difference is that Pebbles defines a Personal Universal Controller (PUC) Specification Language through which the device parameters that can be controlled are specified. The PUC language does not only specify these control parameters, but also a protocol for transmitting changes to the state of these parameters between the appliance and the controller. Essentially, the player in Pebbles has to interpret the PUC specification published by a device in order to generate its interface, i.e. applies an XSLT-like transformation to obtain from the XML representation of the controlling parameters a set of graphical controls. Unfortunately, Pebbles focuses all its work on the presentation and interaction process and has not solved the important service discovery issues that EMI<sup>2</sup>lets has addressed. Moreover, in EMI<sup>2</sup>lets the designer of a smart object is the one who decides which will be the best look and feel of the graphical interface to control the smart object, whereas in Pebbles that decision is left to the player itself.

The Obje software architecture [4] is an interconnection technology that enables digital devices and services to interoperate over both wired and wireless networks – even when they know almost nothing about one another. Their goal is to be able to simply plug new device types into the network and then all existing peers on the network will be able to use them. Similarly to EMI<sup>2</sup>lets, Obje is agnostic to the underlying discovery and communication mechanisms. It also defines four simple abstractions that remain constant and all peers on the network understand: a) connect to another device, b) provide metadata about itself, c) be controlled, and d) provide

references to other devices. In addition, it defines a messaging protocol over TCP/IP that every Obje-enabled device must implement. The main difference between EMI<sup>2</sup>lets and Obje is that whereas in the former the developer of a smart object is the one who decides what the user interface presented to the end user will look like and what functionality it will have access to, in the Obje case the responsibility for determining appropriate interactions shifts from the developer to the end user. In other words, the programming on each device in Obje only tells the device how to interact with peers using the abstract mechanisms previously mentioned. The authors of Obje argue such semantic ignorance is necessary for open-ended interoperability. However, this flexible approach implies that they will need to provide tools that let end-users compose and configure devices within a space. Our approach in EMI<sup>2</sup>lets is much simpler and almost as flexible. The smart object developer decides the best and richest multiplatform (PC, PDA and mobile phone) user interface to control an object. Through EMI<sup>2</sup>lets, the end-user can directly operate with its surrounding objects. As a second drawback, Obje only runs on the PC platform and provides the capability for the end user to integrate different components within a smart space, but does not make the smart objects embedded in AmI spaces readily available for the end-user to control as EMI<sup>2</sup>lets does.

## 7 Conclusion and Future Work

This work has described the design and implementation of a novel reflective middleware, which provides universal active influence capabilities to mobile devices over smart objects, independently of the objects location. This framework presents the following features:

- Transforms mobile devices into universal remote controllers of smart objects.
- Enables both local and global access to those smart objects, i.e. anywhere and at anytime.
- Independent and extensible to the underlying service discovery and interaction, graphical representation and persistence mechanisms.
- Enables AmI using conventional readily-available hardware and software tools.
- Follows a “write once run in any device type” development philosophy.

In future work, we want to add more sophisticated service discovery and context negotiation features between EMI<sup>2</sup>let Players and Servers, following the WebProfiles model described in [20]. In addition, we want to enable the cooperation of smart objects, for instance, through the creation of a distributed shared tuple space. Finally, we intend to incorporate Semantic Web features to our framework, which may move the user “out of the loop” in the EMI<sup>2</sup>lets discovery and execution process, as suggested in [6].

## Acknowledgements

This work has been financed by a SAIOTEK 2004-05 grant from the Basque Government and the Cátedra de Telefónica Móviles at the University of Deusto. This work won the Image Cup 2005 competition organised by Microsoft Spain.

## References

- [1] Bluetooth Specification version 1.1, <http://www.bluetooth.com>, (2005)
- [2] Beigl, M., Gellersen H.W., and Schmidt, A.: MediaCups: Experience with Design and Use of Computer-Augmented Everyday Objects. *Computer Networks, Special Issue on Pervasive Computing*, Vol. 25, No. 4, (2004) 401–409.
- [3] Czerwinski S., Zhao B. et al.: An architecture for a Secure Service Discovery Service. *Proceedings of MobiCom'99*, (1999)
- [4] Edwards W.K., Newman M. W., Sedivy J.Z. and Smith T.F: Bringing Network Effects to Pervasive Spaces. *IEEE Pervasive Computing – Mobile and Ubiquitous Systems*, Vol. 4, No. 1, (2005) 15-17
- [5] Grace P., Blair G. S. and Samuel S.: A Reflective Framework for Discovery and Interaction in Heterogeneous Mobile Environments. *Mobile Computing and Communications Review, ACM SIGMOBILE*, Vol. 9, No. 1, (2005) 2-14
- [6] Lassila O. and Adler M.: Semantic Gadgets: Device and Information Interoperability in Kalle Lyytinen & Yongjin Yoo (eds.): "Ubiquitous Computing Environment", Case Western Reserve University, (2003)
- [7] López de Ipiña, D., Mendonça P. and Hopper A.: TRIP: a Low-cost Vision-based Location System for Ubiquitous Computing, in *Personal and Ubiquitous Computing*, Vol. 6, No. 3, (2002) 206-219
- [8] López de Ipiña D., Vázquez I. and Sainz D.: Interacting with our Environment through Sentient Mobile Phones. *Proceedings of 2<sup>nd</sup> International Workshop in Ubiquitous Computing (IWUC-2005)*, ICEIS 2005, ISBN 972-8865-24-4, (2005) 19-28
- [9] Microsoft Corporation.: Mobile Developer Center, <http://msdn.microsoft.com/mobility/>, (2005)
- [10] Myers B.A.: Using Hand-Held Devices and PCs Together. *Communications of the ACM*, Vol. 44, No. 11, (2001) 34 - 41.
- [11] Rohs M., Zweifel P.: A Conceptual Framework for Camera Phone-based Interaction Techniques. *Pervasive Computing: Third International Conference, PERVASIVE 2005*, Lecture Notes in Computer Science (LNCS) No. 3468, Springer-Verlag, Munich, Germany, (2005)
- [12] Scott D. et al.: Using Visual Tags to Bypass Bluetooth Device Discovery. *ACM Mobile Computing and Communications Review*, Vol.9, No.1, (2005) 41-52.
- [13] Shadbolt N.: Ambient Intelligence. *IEEE Intelligent Systems*, Vol. 2, No.3, (2003)
- [14] Siegemund, F. and Krauer T.: Integrating Handhelds into Environments of Cooperating Smart Everyday Objects. *Proceedings of the 2nd European Symposium on Ambient Intelligence*. Eindhoven, The Netherlands, (2004)
- [15] Sun Microsystems, Inc.: Java 2 Platform, Micro Edition (J2ME), <http://java.sun.com/j2me/> (2005)
- [16] Sun Microsystems, Inc.: Jini Specifications Archive - v2.1, [http://java.sun.com/products/jini/2\\_1index.html](http://java.sun.com/products/jini/2_1index.html), (2005)
- [17] Symbian Ltd.: Symbian OS – the mobile operating System, <http://www.symbian.com/>, (2005)
- [18] The Universal Plug and Play Forum: <http://www.upnp.org/>, (2005)
- [19] Vázquez, J.I., López de Ipiña, D.: An Interaction Model for Passively Influencing the Environment. *Adjunct Proceedings of the 2nd European Symposium on Ambient Intelligence*, Eindhoven, The Netherlands, (2004)
- [20] Vázquez, J.I. and López de Ipiña D.: An HTTP-based Context Negotiation Model for Realizing the User-Aware Web. *1st International Workshop on Innovations In Web Infrastructure (IWI 2005)*, Chiba, Japan (2005)
- [21] Zhu F., Mutka M.W., L.M. Ni.: Service Discovery in Pervasive Computing Environments. *IEEE Pervasive Computing*, Vol. 4, No. 4, (2005) 81-90

# Ambient Interfaces for Elderly People at Home

Fausto J. Sainz Salces, Michael Baskett, David Llewellyn-Jones, and David England

Liverpool John Moores University,

School of Computing & Mathematical Sciences, Byrom Street, Liverpool, L3 3AF  
{cmsfsain, M.Baskett, D.Llewellyn-Jones, D.England}@livjm.ac.uk

**Abstract.** The elderly population in the world is increasing rapidly and consequently so is demand for new technologies that allow them to live independently. Facilitating the control of household appliances and the home environment through various devices that encompass multimodal and ambient interfaces seems a way to achieve this. In this paper, we lay out the theoretical principles relating to the accommodation of technology for use in the home among older people, followed by a report supporting these principles based on experiments we have carried out. Three modalities of output – audio, visual and multimodal – were tested using two different devices – palmtop and laptop – as realistic prototypes of household appliance controllers. Through experimental design, the applicability of using icons and musical earcons as a medium to transmit information to the user and its suitability to the home was investigated. The use of musical earcons allowed the potential for an ambient interface to be compared with a traditional visual interface for older people. Results showed participants performed markedly better using the multimodal and visual interfaces than with the audio interface. In addition, both groups performed better using the palmtop as compared to the laptop.

## 1 Introduction

With the rising numbers of older people, increasing expectations for a better old age, a reduced capacity of families to provide care, and pressures on public finances, high priority is being given by governments around the world to research into human ageing and how to provide a better quality of life for older people. Designing artefacts while considering the requirements of users with extra-ordinary needs, such as the elderly and disabled, would result in a product more widely useful and a space where the disadvantaged do not become even more disadvantaged. New technology can play a very important role in the quality of life of elderly and disabled people who wish to continue to live autonomously and can be assisted by technology in their daily routines. This is achieved not least by helping them regain some control and some very psychologically valuable independence [1]. Telehealth, telecare, telemedicine and personal safety systems are all examples of this trend.

We believe that a multimodal approach facilitates users' adaptation to the format of information displays that suit them best, due to external circumstances or personal preferences. Multimodal interfaces will benefit all users and promote "universal access" [2, 3]. In addition, the use of audio and non-visual interface aspects allows passive and ambient interaction with devices that is hard to achieve through the visual medium.

This paper describes the process used to develop and test a multimodal interface, along with the results obtained through an extensive and rigorous testing process. We begin with an introduction to the issues regarding the elderly population, followed by a section about multimodality as a possible solution to improve interfaces. The next section is then concerned with the area of domotica, technology and living arrangements. We then consider previous research in the area, followed by a section describing the ethics and design process of the experiments. Finally, we describe the experiments conducted and the results obtained from these experiments

## 2 Elderly

The elderly population is an especially big segment of the general populace [4]. In addition, life expectancy in the European Union has increased considerably since 1980, with an increase of almost 5 years for both males and females (Eurostat)<sup>1</sup>. Certain characteristics that define this segment of the population are also the cause of impairments that we can experience at any point during the course of life, and are therefore unexpected *a priori*. The elderly population is a very heterogeneous group with a wide variety of characteristics [5].

Older people and people with special needs are both fast-growing segments of the population, but the group of elderly people is also highly diverse, covering a broad range of different abilities and weakness. It must not be forgotten that we can all be handicapped by our environment, and statistically, we will all, at some point in our lives (as we live longer), become disabled in some way [6].

The boundary of old age is purely a chronological event and bears no relationship with how the individual feels or his/her appearance. Being old, by itself, is of no relevance to computer systems. However, the onset of disabilities and disease is. In relation to new technologies, the problem comes when instead of offering facilities, new systems result in the denial of older users – and not so old users – with a wide range of disabilities, access to functionality, thus hindering their interaction with new solutions and preventing them from using the newly developed technology. In some cases, new technological solutions can worsen the user's previously satisfactory experience with the apparatus or device and consequently may even dampen enthusiasm for new technology in general.

Older people, who are not familiar with computers, will work with computer-based systems if they see the benefit of them. More of them will become familiar with information technology in the coming years and they will demand more of these kinds of services. As shown, by the fact that they represent the fastest growing demographics on the Web [7], older people are a group of users that should be taken into account when designing new technologies.

The aging process is generally related to a decrease in functionality and sensory loss, as well as cognitive slowdown. Changes in cognitive functions during old age affect the speed of information processing and memory. This can, for example, cause problems if time outs in operating procedures are too short.

---

<sup>1</sup> Life expectancy at birth. Males 1980 : 70.5, 2002: 75.5 Females 1980: 77.2, 2002: 81.6. Eurostat.

## 2.1 Ageing Population

The older population of the world (developed world mainly) is increasing rapidly. Data from the United Nations suggests that this trend is likely to continue through the coming centuries<sup>2,3,4,5,6,7</sup>.

“Over the next quarter century Europe is projected to retain its title of ‘oldest’ region in the world. Currently older people represent around 20% of the total population now and will represent 25% by 2020.” [8].

Efforts should therefore be directed towards helping the increasing numbers of older people (our fastest growing segment of the population) to enjoy a good standard of life.

Life expectancy in the European Union has also increased considerably since 1980, with an increase of almost 5 years for both males and females (Eurostat)<sup>8</sup>. As pointed out by Donnellan<sup>9</sup> [8]:

*“In 1997, life expectancy at birth in the UK is expected to be 74 years for men and nearly 80 years for women. Life expectancy increases with age: in England, a man 60 can expect to live a further 18.7 years to 78.7 and a woman aged 60 a further 22.6 years to 82.6. Life expectancy in the U.K. has increased steadily though this century.”*

In many countries the very old – those over 75 – are the fastest growing portion of the older population. However, as predicted by the WHO [9] it is not only Europe that is experiencing a demographic transition from high mortality/high fertility to low mortality/low fertility pattern; it is a global pattern<sup>10</sup>.

### 2.1.1 Ageing Effects

When we refer to the older population, we are taking the most widely accepted definition that refers to those people over the age of retirement: from 65 years of age upward. With Tinker [10] we are aware of the differences among the older people population that make it difficult to generalise.

There are peculiarities of this age group that made the design of objects, in general, and computer products, in particular, a rather challenging experience. Characteristics

<sup>2</sup> “One in every ten persons is now 60 years or above; by 2050, one out of five will be 60 years or older; and by 2150, one out of three persons will be 60 years or older.”

<sup>3</sup> The older population itself is ageing. The increase in the number of very old people (aged 80+ years) is projected to grow by a factor of 8 to 10 times on the global scale, between 1950 and 2050.

<sup>4</sup> The majority of older persons (55%) are women. Among the oldest old (80 years or older), 65 per cent are women.

<sup>5</sup> One out of five Europeans is 60 years or older.

<sup>6</sup> In some developed countries today, the elderly proportion represent close to one in five.

<sup>7</sup> At the individual level, it is estimated more than 20 years will be added to the average life of an individual by the end of this century.” (United Nations/ Division for social policy and development).

<sup>8</sup> Life expectancy at birth. Males 1980 : 70.5, 2002: 75.5 Females 1980: 77.2, 2002: 81.6. Eurostat.

<sup>9</sup> (Social Trends 27, 1997, 1997, chart 7.1). Health and Personal Social Services Statistics 1998, tables A1 and A2.

<sup>10</sup> It is estimated that by 2050, 16% of the projected 9 billion person global population will be older than 65 years.

such as different degrees of disabilities and impairments [11] can make it an almost impossible task, as can the varying degrees of cognitive functioning that challenge the ways in which to promote easy learning and defy the development of custom interfaces and artefacts for them. As pointed out by Hawthorn:

*“the effects of age become noticeable from the mid forties onward. So this is not just about design for yet another minority group, the one termed senior citizens. This paper is concerned with design for the second half of our lives and for a group that will shortly be nearly half the workforce and over half of the adult population.”* [5]

This shows how important it is to think about the older population in terms of design. It is also important to remember that some older people have just minor disabilities and/or a combination of them and other factors that affect their lives in such a way that they become housebound [12]. We will discuss these in the following section.

### 2.1.2 Sensory Loss

The aging process is generally associated with a decrease in functionality and sensory loss, as well as cognitive slowdown. One of the most obvious effects of ageing is sensory loss [5], [13], [14]. For many older people, some abilities tend to remain intact or slightly affected, while others suffer from a great deal of loss. This depends greatly on the individual.

The decrease in functional ability can affect the following adversely.

- Sensory organs (vision, hearing, smell, tactile sensation, taste, *etc.*).
- The information process capacity.
- Speech intelligibility.
- Reduced speed and increased variance in the timing of precise movements.
- Length of time required to retrieve information from memory, *etc.* [15].

Sensory loss does not occur as a homogeneous process and individual variations can be very significant.

The combination of audio and visual displays in an interface could minimize the effect of any sensory loss experienced by the user and help in getting all the information to the subject. People who have acquired a sensory impairment learn to obtain information from a different sense [12], and the discovery of multisensory cells helps to corroborate the idea that we experience the world in a multisensory manner [16]. Whether loss of vision arises through ageing, accidents, or any other reason, the use of sound can help older people (and other users) to adapt to these new conditions so they can maintain a normal, active lifestyle. By means of the multimodal interface, the effects of disability can be compensated through the use of other modes to convey information.

### 2.1.3 Vision Loss

Although vision loss is not the only problem affecting older people, it is reasonable to think that the more varied the stimuli that can be delivered, using different sensory channels, the more effective the communication of information to the user will be.

Using various channels would seem to be better than just one. Although it is interesting to note that older adults modify their behaviour in a manner appropriate to their changing abilities.

Changes in visual perception, (from changes in visual acuity, accommodation, contrast acuity, depth perception, to visual acuity) are also noticeable from early adulthood as reported by the INCLUDE project [14]. Letter acuity declines swiftly from 50 to 90 years of age. Ability to see in dusk and colour vision decreases with age. The same happens to accommodation: the decrease in adaptation to see near-distances, a problem that sometimes starts as early as the age of 20, and the lengthening of time required for adaptation to see in poor lighting conditions. Older adults had decreasing contrast sensitivity at higher spatial frequencies. The contrast sensitivity for old people is also 1.5 to 4 times lower. From 30 to 80 years, depth perception also becomes poorer, with an increase in threshold from 100 to 300 sec per arc [14].

As an indication of the widespread decline in visual acuity among ageing humans, corrective lenses are almost universally needed for good visual acuity at optical distances inferior to 0.5 meters from around the age of 40 onwards [13]. Again, this supports the idea that design for older people not only will satisfy their needs, but also those of a wide range of the population; indeed anyone that suffers any kind of difficulty in using technology due to the fact that they are impaired, or artificially impaired by circumstances.

Some visual capacities tend to decrease with age, either due to the effects of aging or the occurrence of diseases more prevalent with age. Cataracts, glaucoma, age-related macular degeneration and diabetes retinopathy are other aetiologies that will develop with age<sup>11</sup> [17]. It is also known that the vision of older adults is much less sensitive to fine-grained and moving targets than is the vision of younger adults.

#### 2.1.4 Hearing Loss

The ear is no exception to the gradual degeneration of body tissues that comes with aging. The degree of hearing loss though is not a standard process and there are considerable differences in the capacity loss from person to person, from complete deafness, in which no sound is heard, to a slight difficulty in following speech in a conversation taking place in a noisy situation.

The hearing sense is affected in several ways by the ageing process, from the ability to localize a sounds source, deafness and hardness of hearing to sensitivity loss. Some of these degenerative processes – such as presbycusis (loss of sensitivity) – start as early in life as approximately 25 years of age.

Problems in hearing capabilities are familiar to all of us when dealing with older people. Fozard *et al.* [13] reported an increase in thresholds for pure tones of high frequencies. These increase regularly at a rate of about 0.3 db a year through to the age of 60, but from 80 to 95 years of age the decline rapidly increases to 1.4 dB per year.

---

<sup>11</sup> Blindness and impaired vision become more common as we get older. In the UK 42% of people over 75 will develop cataracts, almost 50% will have age-related macular degeneration and 7% will be affected by the most common form of glaucoma. Significant numbers of people will suffer from more than one condition. <http://www.ageing.org/research/why.html#top>



To give an idea of the effect of such deficit over the long term, Table 1 below shows the levels of hearing loss in terms of decibel thresholds.

We hear across a wide range of frequencies or pitches. The human ear is capable of detecting sound waves with a wide range of frequencies, ranging between approximately 20 Hz to 20 000 Hz. A high pitch sound corresponds to a high frequency and a low pitch sound corresponds to a low frequency. Thus 250 Hz is a low frequency sound and 10000 Hz a high frequency sound. Pitch is our perception of the rate, or frequency, of the vibrations that make up the sound waves. Faster vibrations create higher pitches and slower vibrations create lower pitches.

**Table 1.** Hearing loss in terms of decibels

| Hearing threshold (in decibels, dB) | Degree of hearing loss | Ability to hear speech  |
|-------------------------------------|------------------------|---|
| 0–25 dB                             | none                   | no significant difficulty   |
| 26–40 dB                            | mild                   | difficulty with faint or distant speech                                   |
| 41–55 dB                            | moderate               | difficulty with conversational speech                                     |
| 56–70 dB                            | moderate to severe     | speech must be loud; difficulty with group conversation                   |
| 71–90 dB                            | severe                 | difficulty with loud speech; understands only shouted or amplified speech |
| 91+ dB                              | profound               | may not understand amplified speech                                       |

At around the age of 60, identification of pure tones in the low frequencies is within normal margins, but for high frequencies, hearing loss is in the range of mid to moderate. With ageing the thresholds also increase, up to the age of 90, thus resulting in mid hearing loss in the low frequencies and moderate to severe hearing loss in the higher frequencies [13]. Men, on average, have poorer auditory thresholds than women.

The overall trend is for the high tones to be heard gradually less well than the low ones, as well as the gender differences as noted above [18].

For example, the sound produced by a drum or a deep male voice, and vowel sounds, are low-pitched, while notes on a flute, a child's voice and some consonant sounds such as 'f', 's', 'sh' and 't' are high-pitched. Although pitch is directly related to frequency, they are not the same; frequency is a physical phenomenon whilst pitch is a perceptual one (psychoacoustic, cognitive or psychophysical).

Lysons [19] also considers the importance of exposure to noise in relation to presbycusis, as this is more pronounced in some people than others. He cites Rosen, who studied the Maabaans:

*“a primitive African tribe living in an environment where the noise levels only occasionally exceeded 40db, indicated that a male Maabaan aged 70 and 79 had keener hearing than a sample of Americans in the 30-39 age range who had been exposed to the noises of modern civilization”.*

Since much of society will be constantly surrounded by noise levels well above those experienced by the Maabaan's, presbycusis seems to be unavoidable for the vast majority of our population. Lysons [19] also points to age related loss:

*“It has been estimated that about 25 per cent of 65 years old have more difficulty hearing than they did at 30. At 70 and 80 the percentage increases to 33.3 and 50 respectively.”*

Regardless of the aetiology of the condition, whether conductive or sensory-neuronal – losses can be due to a wide variety of causes – the occurrence of it in old age cannot be neglected.

### **2.1.5 Tactile Sense**

The decline in sensory perception is also noticeable in the tactile sense. Stevens *et al.* [20] study showed that tactile acuity thresholds averaged from 22% to 80% higher in older subjects than in younger ones. If there is an associated disease, the problem is markedly exacerbated, as demonstrated by Sathian *et al.* [21] in consideration of patients with Parkinson's disease. They showed a twofold increase in the tactile spatial threshold. These patients were also impaired in tactually discriminating grating roughness, with differences over three times higher than a group of older people without Parkinson's disease. Older subjects also tested consistently worse in ability to discriminate tactile gaps, orientation of lines and length of lines as reported by Stevens.

#### *Implications for design*

These characteristics of the human sensory systems should determine the parameters for the design of any interface aimed at the older population and, by extension, at the general public.

*“The ability to localize a sound source is dependent primarily upon the ability of the auditory system to process acoustic information on time and intensity. Localization of low frequencies is determined mainly by small differences in the arrival time of the sound at the two ears (for high frequencies, localization depends mainly on the inter-aural intensity difference). Men over 60 years of age require a greater time delay across ears for accurate localization of low frequencies, but there is no age effect in localization of high frequencies. Sound localization by the older people is poor when the precedence effect between ears is 0.5msec or less.”* [14]

Again, the use of technology can diminish the effects of sensory loss. Considerations on the effect of aging in human decline [22] [23] – such as sensory, cognitive and physical decline – must be taken into account, *e.g.* reduced sensitivity to colour, loss in the ability to detect tones over all frequencies, memory loss and attention deficits. A number of previous research recommendations can be followed in relation to this [24] [25]. An immediate direct consequence of this approach is that it will extend the time users are able to use applications.

### 2.1.6 Cognitive Decline

On top of the sensory deterioration the body is subjected to, there are also certain cognitive problems acquired with aging. These deficiencies have effects on the processing of information, such as slow response time and the speed with which information is processed, *etc.* [23].

*“An older user typically has: limited short-term memory; lower co-ordination capacity; lower sensory capability, and slower ability to react.”* [26].

Changes in cognitive functions in old age affect the speed of information processing and memory that can cause problems if time outs in operating procedures are too short.

It is thus agreed that concentration only on a sensory solution will accommodate just 8.5% of users [26] as most older people have multiple capability losses. In designing an interface, it is therefore important to take into account not just the sensory deficiencies, but also the cognitive and motor deficiencies. In the design process discussed later on, we have therefore taken a holistic approach in order not to invalidate the research on musical earcons that was the objective.

## 2.2 Psychological and Social Gain

New technologies can augment the bodily functions [27] and enable users to interact with their surroundings in new ways. They can also make disabled users amongst the major beneficiaries of these new advances providing empowerment, independence and also alternatives. It is important that these new developments are not perceived as marginalising or stigmatizing, but rather as “normalizing” tools. Stigmatizing and addressing single disabilities, as pointed out by Hansen [28], are two of the problems assistive technologies still present nowadays.

Consider some of the additional benefits that can be gained through the inclusion of assistive technologies in the home, that allow a person additional freedoms to live in a more independent manner. Regardless of the level of complexity embedded in to the house, the possibilities offered by this technology are considered essential for some individuals’ happiness and well being [29]. People with disabilities are potentially the major beneficiaries of new technological advances. We have to make sure, when using new technologies in design, they are empowering and satisfying users and not disabling and discriminating against them in any way. The interest developed recently in multimedia can be of great help to these less fortunate citizens. People with certain dependencies on others can attain that independence with the use of assistive or standard well designed enhanced technology that allows them to live in a house that is supportive to their individual needs. It is important to see the social gains of this [30], as well as the individual psychological ones. There are significant psychological gains when people stay, for as long as possible, in the dwelling perceived by them as their home, where objects surroundings and the environment interact to give the person a strong self identity [31, 32].

In the care of elderly people, there is a trend to move the facilities rather than the people, so that individuals can remain in their own homes, encouraging self-management, autonomy and independence.

The UN principles for older persons encourage the idea of independent living for the elderly, as shown in points 5 and 6:

*“5. Older persons should be able to live in environments that are safe and adaptable to personal preferences and changing capacities.*

*6. Older persons should be able to reside at home for as long as possible.” [8].*

Extending the time people can stay in their homes is highly valued by the older population. To turn the longing desire of independent living in old age into a reality, it is necessary to implement technical solutions that enhance the safety and quality of life available to older people. A household control system that ensures users will not sustain injury, harm or undue risks, where elders can feel competent as well as feel good, is one step forward into the environmental improvement needed to achieve independence for as long as possible. Whilst some residential and care homes for elderly people are welcoming and residents feel at home, spending the last years of their lives in a warm environment, other institutions continue to appear more as a place for patients to die than a truly humane place for people to live [33].

The help that technology can procure to those more vulnerable in society is not limited to palliate physical impairments and/or disabilities, but also to ameliorate the effects of cognitive deficiencies among the population. Certain groups can benefit from new developments in science and technology that will keep them independent for longer [34]. Memory aids and reminiscence exercises [4] are examples of these possibilities.

As previously stated [35] [36], with assistive technology, older people can remain longer in their own homes, thus saving money on expensive state-provided care. These new technologies can lessen the risks of accidental physical damage existing in the home environment. Not only that; these new technological solutions can also provide new answers to old problems and have an impact in an area of daily living of great concern to the older population: safety. Amongst older people, safety is one of the most important concerns [35].

It is safe to say that the use of technology can promote psychological well-being via improvements in independence and ameliorating the effects of cognitive and physical disabilities in various ways. On the other hand, there is a risk that individuals living independently might also isolate themselves from social interaction due to the easy use of technology that facilitates contact with other people, or having social contact only through mediated technology (internet, telephone, *etc.*). The substitution of human contact by the use of technology could also lead to new problems.

In the preceding sections, we have considered how the population is ageing worldwide, focusing on the U.K. and Europe, and how that ageing process affects the different senses, as well as cognitive capabilities. We have also seen how the ageing process affects the way in which people interact with their environment and how it can have an impact on their daily lives. We have exposed how the consequences of the ageing process should influence the design process of products that are aimed at the older population and also those aimed at the general population. The importance of helping people to live independently for as long as possible – and how this achievement can help the individuals, as well as the society in which they live – has also been reflected on.

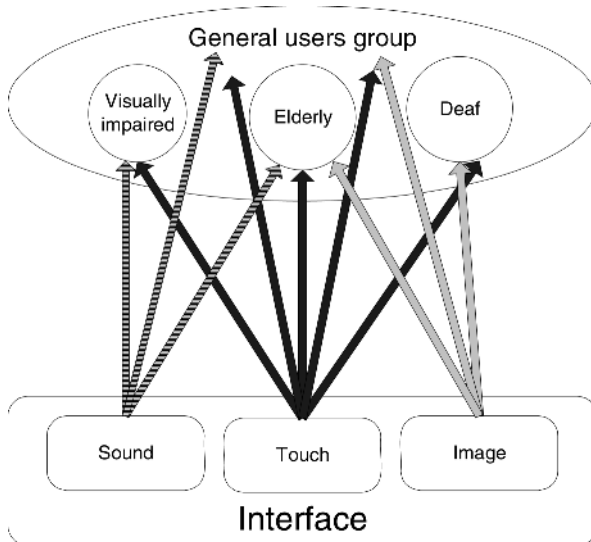
### 3 Multimodality and Ambient Interfaces as a Solution

At the foundation of any project that intends to develop products accessible for the elderly, considerations concerning the effect of aging in human decline [22] [23], such as sensory, cognitive and physical decline, have to be taken into account. The introduction of different modalities into the interface can facilitate the transmission of information to users. This is especially appropriate in the case of elderly and disabled users, who can be counted amongst the major beneficiaries of these new advances. Such advances provide empowerment, independence and also alternatives. It is important that these new developments are not perceived as marginalising or stigmatizing, but rather as “normalizing” tools.

The benefits associated with multimodal interfaces apply particularly well to older people, due to the ambient nature that auditory interfaces allow. The use of auditory earcons as a method for the transmission of information means that users do not have to be concentrating directly on the interface in order to receive and understand such information.

As previously suggested by Sainz *et al.* [37], multimodal interfaces can improve user interaction with a household appliance controller. We can take advantage of the different senses to transmit information about the household environment.

The use of multimedia and especially sound is extremely important to reach certain groups of users; it is an essential information retrieval option for visually impaired users, and the auditory channel offers the advantages of a ubiquitous approach [38, 39].



**Fig. 1.** Effects of a multimodal approach

The advantages of using multimedia interfaces are clearly seen in Arroyo *et al.*'s experiment [40] on interruptions. It also shows the importance of individual differences among people when reacting to different kinds of stimuli. Fraser *et al.* [41] found that people interacting with a graphical interface were able to select targets

significantly faster when they were given targeting feedback, this being either audio or visual. They also found that participants made fewer errors. Redundancy thus can improve the usability of graphical user interfaces.

The richness provided by multimedia and multimodal interactive systems is especially appropriate for older people who often suffer from reduced sensory, motor and intellectual capabilities [4]. Particularly important to the group of older people, where cognitive capabilities are diminished, is the fact that the multimodal interface gives people the “feeling of experiencing information instead of acquiring it” [42], due to the increased stimulation of the senses, strong recognition effects and higher emotion arousal. Thus, it is believed that carefully applied multimodality improves user-friendliness, the impact of the message, the entertaining value of the system and improves learning of the system.

Depending on the task to perform and the environment in which the task is set, audio or visual displays can show superiority. Fitch multivariate task, where both task and stimuli were complex, audio displays elicited faster and more accurate responses than the visual display. Often a bi-modal or multimodal presentation could benefit from the advantages of both audio and visual modes, making it the ideal option.

A multimodal approach allows the delivery of information through two or three channels to a wider segment of the general population, as shown in Fig. 1.

### 3.1 Technology

In recent years there has been a noticeable increase in the use of computers in everyday life. There is a growing presence of computer systems all around us, from domestic environments to work settings and even in the street. Personal computers are becoming part of daily life [6], whether we like it or not, and people are becoming increasingly dependent on them [43]. Computer systems can help in many tasks such as the monitoring of surroundings, as well as other activities that are not directly related to the office environment. Although computers offer a wide range of possibilities, their facilities must also be adapted to a very varied population, where small groups with different physical, sensory and cognitive capabilities either already make use of them or would use them if it were possible for them to gain access to their applications and functionality.

Differences in capabilities amongst the general population – whether they are physical, sensory or cognitive – impose a new approach to the design of computerised applications. Such new approaches are necessary to ensure access is gained by as many users as possible. The potential for technologically aiding those less physically and mentally favoured is significant and promising. Computers can be put to use in improving the quality of life for people with dementia by providing users additional communication possibilities, leisure or information seeking, especially through the Internet [44]. The communication possibilities offered by the computer can facilitate interactions with relatives, carers, and help develop new skills and hobbies, if the interface doesn't act as a barrier to people with cognitive difficulties.

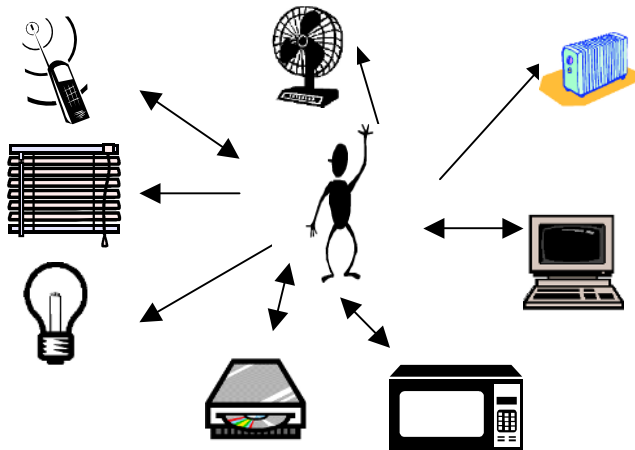
As discussed in earlier sections, new technology can also play a very important part in the quality of life of older people and disabled people who wish to continue to live autonomously and can be assisted by technology in daily routines by helping them regain some control and some psychologically very valuable independence [1].

There can also be a constructive use of technology to combat social exclusion of the less favoured groups as done by projects such as ACTSLAINE, or Listen [45] favouring better levels of interactivity for these groups. It is important to notice that the group of older people and disabled users are probably “in more need of modern technologies than other.” [43]. Another project (Equality), aimed specifically at older people and disabled people, shows the importance and awareness of providing services that will improve people’s equality [46]. Other projects such as EMBASSI [47] attempt to tackle the framework of support for users with special needs in their interaction with public information systems.

Due to the fact that in developed countries societies are ageing, as is the proportion of people with disabilities and chronic diseases – that in one way or another impede normal living – the emphasis in clinical practice is shifting towards maintenance and palliation. This can be done with the help of computer systems that allow citizens to remain at home for longer periods of time without the need to move into institutional care.

## 4 Domotica

People accept computer based technology and even computers nowadays as omnipresent artefacts. This has been helped by the fact that in many areas of the public domain, such as banks, airports, shopping malls and bus terminals, it is not uncommon for activities to be performed through computer-based technology. The predictions made by Venkatesh [48] about household automated tasks, smart appliances and smart homes appear to be becoming reality. Heating systems, blinds, fans, and other appliances (see Figure 2) can be programmed and operated via a shared joint interface.



**Fig. 2.** Possible household appliances controlled through a common interface

Difficulties in operating household appliances as well as performing other daily chores are part of the effects of, in many cases, growing older sensory and cognitively. Aside from the economic factors and motivations, recent efforts in the

development of hardware and software solutions to these problems [49] through the development of what is called assistive technology, proves the importance of enhancing the quality of life of the older people and the disabled. Most of these solutions have a positive emotional impact on their users, thus improving the quality of life, not just by making life easier and more pleasant in terms of the object being manipulated and interaction with the physical surroundings, but also by indirectly promoting the disappearance of physical constraints and thus having a very positive mental impact on users.

There has been an increase in the research undertaken in the field of domotica over the last decade. The increase on feedback options, as well as operating methods, allows the public a much wider range of options than a few years ago. Voice activated environment controls and other sophisticated control panels are examples of the effort put into making technology for the home increasingly acceptable, non-cumbersome and desirable. Ifukube [50] suggested that intelligent tools that are able to recognize both verbal and non-verbal information should be developed. He goes further, suggesting that these tools should be able to reproduce non-verbal as well as verbal information provided by text or speech and that they should also include actuators that comprise “artificial hands, fingers and a face.”

These advantages also stir more demand from users, which, in turn, fosters the development of more products. The introduction and rapid development of new technological solutions, such as the microprocessor controllers found in most domestic appliances nowadays, new complex digital technologies and digital automation facilitate the communication between appliances with control devices [51] with other appliances and also with users. Such technological advances promote a rapid development of home safety and comfort.

Domotica is an area of research that is growing fast and can benefit from this research. Recent developments in home networking are having a great effect on the quality of lives of disabled and older people. The home of the future, and in some advanced cases, of today, embraces technology as an integral part and a fundamental concept. Nowadays we have the means to manage almost anything that uses electricity and to adapt devices to be controlled electronically (doors, blinds, baths, beds, lights...). In the not so distant future, digitally engineered domestic life will provide us with commodities and services that will make home life a better experience. Ubiquitous and pervasive computer systems will be used in tomorrow's home infrastructure to provide services that are aware of users needs [52].

The use of technology in the home is on the increase. Smart homes, telecare, telehealth [53] and telemedicine are examples of the practical application of technology in the home environment. All of these examples are oriented to allow people to remain in their homes with a good quality of life. There are a large number of technological solutions available for older and especially disabled people. These range from solutions to very specific problems (*e.g.* voice reproduction) to more comprehensive systems.

As with other studies, Mann *et al.* [54] found that frail older people experience functional decline over time, as expected. However, they also found that the rate of decline could be slowed, with a consequential reduction in institutional and certain in-home personnel costs, through the introduction of assistive technology and environmental interventions. In a previous study, they also found that increased use of



assistive technology correlated with greater functional independence. It is also suggested that the reduction of cost could be related to prevention of injuries due to the use of assistive technology and environmental interventions.

There are several reasons to encourage the development of smart home and assistive technologies or, in general, the progress on domotica. Among those pointed out by Bonner [55] are improving quality of life, as well as broadening the interest in smart technologies. Systems such as the one proposed by Chan *et al.* [35] could have increased functionality (location tracker, mobility plotter, comfort monitor, *etc.*) and accoutrements that would dramatically improve users' daily lives. The possibilities and combinations offered by several advances in various areas of research (Artificial Intelligence, Security, Neural Networks, Communication systems, Data Analysis, Robotics, Automatization) and technology (hardware and software development) will facilitate, in the near future, a more comprehensive and natural approach to the area of domotica, assistive technology and care in the community [56].

Contrary to common belief, the number of people who live independently is very significant:

*“The vast majority of people over pensionable age live in private (that is, non institutional) housing. Only approximately 5% of people in this age group live in a residential or nursing home. Of those living in private housing, more than a third (39%) live on their own. Just under half, 48%, live just with their partner and 13% live with other people such as a son, daughter or siblings”*<sup>12</sup>.

It is agreed that over three-quarters of the older population in the UK live either on their own or just with their partner in private housing.

Disabled and older people can find problems with some of the appliances used in the house. Doors and windows, for example, are a very common feature in buildings and they are constantly used. The fact that some research has been going on into how to help users manipulate them is a sign of their importance and possible significance at the time of design implementation. Their environment also offers possibilities for modification as the general appliances were not chosen specifically to target a minority of the general population, but new appliances can be incorporated for those minorities such as the physically disabled with specific needs (*e.g.* hoist). As pointed out by Dewsbury *et al.* [27], users could adapt the device in addition to having a say in its design as a limited range of applications could be added or removed. There are some household appliances already on the market, which proves the relevance of the application such as the Remote Home Controller (RHOC by Motorola).

As in other studies, we choose components of the home that are highly likely to benefit from adaptation for use by older people and the disabled when integrated in a computer controller artefact.

There are also prospective reasons to invest time in this area of research for allowing people to continue living independently in their own homes, performing basic activities of daily living such as cooking, bathing *etc.*, instead of having to move – already identified as a deed avoided by older people if possible [10] – to institutional care, which has extensive negative effects for elderly people. Among

---

<sup>12</sup> General Household Survey 1996, table 8.43.

these negative effects were found depression, a feeling of increased isolation and reduced motivation for self care [30].

As pointed out by van Berlo [57], there are several requirements a house for the elderly should provide. These requirements encompass safety and security, comfort, communication and energy saving. They were established by the prospective users of a hypothetical house, and salient among them is the idea of safety and security [58]. Also important to note, and particularly relevant to this study, is the fact that older people themselves required a simple interface.

The capability of our hearing system that allows us to register sound signals incoming from any direction in space without the need to point or direct our external receiver to the emitting source seems to be ideal in the particular case of the household control appliance [59], and HACOP in particular, where we can get information about the appliance that perhaps is not visible at other times (*e.g.* listening to the earcon that represents *the front door is open* while reading a newspaper).

The increase in feedback options, as well as operating methods, allows the public a much wider range of options than a few years ago. These advantages also motivate increased demand from users, which in turn fosters the development of more products. The introduction and rapid development of new technological solutions, such as the microprocessor controllers now found in many domestic appliances are of particular benefit [51].

#### 4.1 Previous Research in the Area

Previous research includes the development of an interface for smart homes based on gestures. This proves how relevant the addition of other sense modalities can be for making the interaction between the user and the interface more naturalistic, *e.g.* ARGUS by Markus Kohler [60]. Vallés *et al.* [61] investigated the possibilities of a multimodal environmental control system for elderly and disabled people that used a touch-screen and an input/output voice and audio module.

Companies such as Fagor, Sincere Kourien, TELETASK<sup>13</sup> and Matshusita, and projects such as Senior Watch, UTOPIA, CIRCA, Equality, KommAS and Memojog also focus on the development of different technologies, techniques and products that will be of great benefit for the elderly and/or disabled. The ultimate goal will be the development of homes that are not specially designed for elderly or disabled people, but homes that can accommodate their needs as time goes by, as with the Joensuu/Marjala development in Finland [34] or the new developing housing state Luis Labin in Spain<sup>14</sup>.

In contrast to the research described in this paper, none of the aforementioned projects investigated the use of a multimodal interface utilising musical earcons to transmit information about household appliance status. Musical earcons allow information to be imparted in a truly ambient and passive manner, since the user can be notified about a change in appliance status without the need for them to actively focus on the interface. This is further enforced by the tonal nature of the earcons

---

<sup>13</sup> <http://www.teletask.be/>

<sup>14</sup> <http://www.fagor.com/es/noticias/index.html>

which sets them apart from other auditory techniques such as speech, which again require increased concentration on the part of the user to interpret and understand the information being imparted. Accordingly, the current investigation fulfils a necessity to clarify the possibilities of delivering information through musical earcons to a resident in a household environment.

#### *Older people's safety and control devices*

There have been a number of previous attempts to apply technology for the safety of the elderly such as Chan *et al.* [62] where a remote monitoring system is proposed, or the KommAS a communication tool for older people with aphasia by Bühler *et al.* [63].

We can also see some awareness of the needs of older people and disabled people in projects such as Equality (UR1010): “where the quality of life of the less favoured European citizens in urban areas can be improved”. In this project, the partners have developed 21 services in different areas of everyday life such as safety devices, tele shopping, *etc.* (TIDE).

Matshusita E. I. Co. is another company that has shown an interest in developing devices that help make older people's lives easier. The company has an experimental home near Neyagawa [64] where computers play a very important role in the care of older people; weight monitors in beds, sensors in the ceiling, *etc.*

There is no evidence that older people are particularly averse to using new technologies. In fact, according to Czaja *et al.* [65], older people are willing, and able, to use computer systems, provided the system is easy. Moreover, according to Bromswell *et al.* [66] whilst the new generations of older people are not technophobes, they would like new technologies not to be imposed upon them. In our study, older participants were eager to learn and experiment with the equipment. The Internet demographics show the rapid expansion of this group of users. There are also economic and market imperatives for the advantages of communication and information technologies to be extended to support this group of users.

One study in domotica, that has showed the growing importance of a multimodal approach was undertaken by Shao *et al.* [67]. They also explored the use of other modalities such as touch or hand gestures as well as voice.

There has also been other research in the area of domotica and older people, as shown in the Sincere Kourien [64] home in Japan, benefit directly from the use of technology aimed at their needs.

As mentioned above, UTOPIA, CIRCA and Memojog are also projects focussing on technologies of benefit to older people and/or the disabled. In the Memojog Project, the aim is providing substantial support to memory, with the aim of reducing the levels of memory normally required to operate computer systems. The system here exposed could help in the exercise of cognitive capabilities, such as memory, recall, decision-making, *etc.*

Lines [68] also investigated the use of speech as spoken dialogue in an intelligent home system as the output media. She reasoned that it may be considered appropriate for the communication of situations of negative consequence within a context in which the user may be conducting hand/eye busy tasks.

BrookesTalk [69] is a web browser that uses speech output for visually impaired people. Designers built a speaking front end onto BrookesTalk after realising the

effects of memory loss and visual impairment, common among older people, can be minimized by speech, thus facilitating interaction on the Web. A new version of BrokesTalk [70] that included non-speech sounds was preferred over the non-sonified version. In the experiment, it was found that the sonically enhanced buttons significantly improved performance (average improvement 8.98%) navigating a web page and participants also found the sonification helpful.

Projects, such as MOSAIC-HS, aim to promote a common standard development for home systems applications, show the importance of this area of research.

The aim of the HACOP system is to achieve maximum personal security avoiding life threatening situations (gas explosion, burglary, *etc.*) thus gaining a feel of security in the house. In addition to this, it also seeks to attain a greater level of comfort provided by new technology at the service of the resident of the dwelling.

Projects such as Senior Watch are aware of the insufficient empirical data about the needs of older citizens that could be met by new information technologies. This project addresses the reality of an increased number of electronic and computerised applications that serve target users such as older or disabled people, developing a methodological approach incorporating the requirement to better understand the needs of older users.

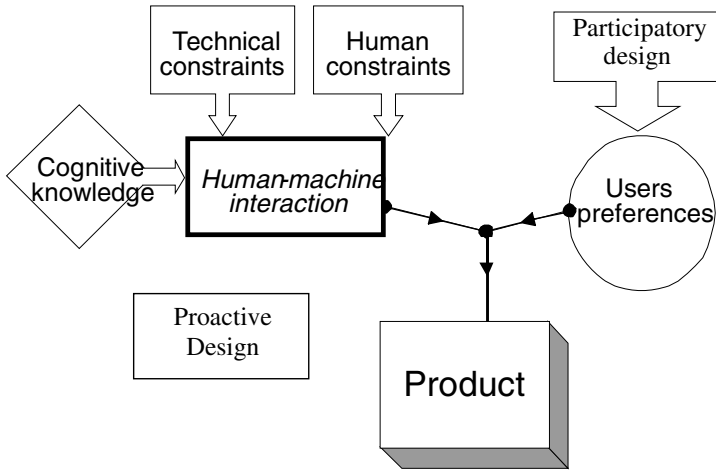
Nowadays the technology needed for the implementation of a remote control for household appliances is readily available [71]. The household appliance field seems to be a good area to test the use of sound.

## 5 Design Issues

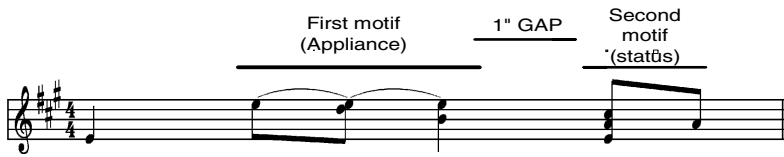
Varying degrees of disability and impairment in the general population, and particularly in the elderly, make it impossible to develop an interface that can be used by all. Moreover, although each individual has particular needs, it is also impractical to develop products for individual consumption. Thus we can only attempt to satisfy a majority of the elderly population, knowing that some of them will be left out. Many products that exemplify inclusive design are highly successful in the market place, as pointed by Arnold *et al.* [7]. Some of these products were designed at the outset to help disabled users, but were quickly adopted by a much wider range of the population. Examples of them include text-messaging facilities, curb cuts and the OXO Good Grips kitchen range.

### 5.1 Design Techniques

There are many techniques developed to help in the design of devices for people with impairments and disabilities. There are, as well, even more to be applied to the design focused on the able bodied or general population. All of these techniques and approaches have their own advantages and inconveniences that prevent them from becoming a definite design strategy that can be used under any circumstance and for any design [72]. The limitations of each approach come from different aspects of the processes, from target users to methodology. To gain as much as possible from the prospective users and knowledge about the subject, it is recommended to use a mixture of techniques that will maximize the resources available to a project. The techniques used should be those that most appropriately fit into the research objectives and sources available.



**Fig. 3.** Design Process. Proposed human and technical factors and techniques involved in the design of a device/application.



**Fig. 4.** Earcon composition

As seen in Fig. 3, we introduced various methods including concepts from cognitive knowledge (human constraints, cognitive knowledge), proactive design (background) and participatory design (user evaluation) in order to achieve the best results and avoiding each method's weaknesses and making the most of its strengths.

The design of the earcons (Fig. 4) and icons followed recommendations from previous studies to make sure they were as suitable for the elderly population as possible.

The use of musical tunes for the audio component of the interface was chosen for its dissimilarities with alarms and emergency calls. It is a more sophisticated method of delivering information in an "elegant" and more expressive form, more suited to an ambient interface design.

## 6 Experiments

The motivation for the research came from the necessity of testing the possibilities of three different interface modes and also of testing the possibilities of musical earcons as an option for the audio interface with ambient properties. To test the possibilities of a multimodal interface in use in the home, we designed an experiment that compared the possibilities of three different interface modalities (audio, visual and multimodal) implemented on two different devices (palmtop and laptop). We also wanted to

consider the potential for the use of musical earcons as a specific method for audio output. The environment chosen was the home environment, as this is an important area in which to concentrate efforts given that older people and the general population can directly benefit from research in the area. The comparison will shed light onto the most appropriate mode of interaction for older people when dealing with remote home control appliances.

For each individual, experiments were conducted over two consecutive days. On the initial day, participants undertook a training session (described in more detail later on) followed by a measured experiment in which their ability to use a simple interface involving four distinct household appliances was tested. On the second day, a further experiment was conducted similarly, using a more complex interface involving an additional two appliances. The characteristics of the original four appliances were left unchanged when using this second, more complex, interface.

Research from the literature, along with previous studies [73] and observations led to the following hypotheses:

Participants will perform equally using any of the three output modalities (audio, visual and multimodal).

Participants will perform equally during the first and second day. The reasoning behind this hypothesis is that as participants were introduced to two new representations of appliances (two new icons and earcons) on the second day they would learn them without difficulty and thus perform similarly to the previous day. This belief arose from Czaja *et al.*'s findings [65] about older people's willingness to use computer systems if, among other conditions, features are added in an incremental fashion.

There will not be any difference in performance between the elder and younger participants.

Participants will perform equally independent of the device used. That is to say that response times for the different modalities will not be affected by the device used in the experiment.

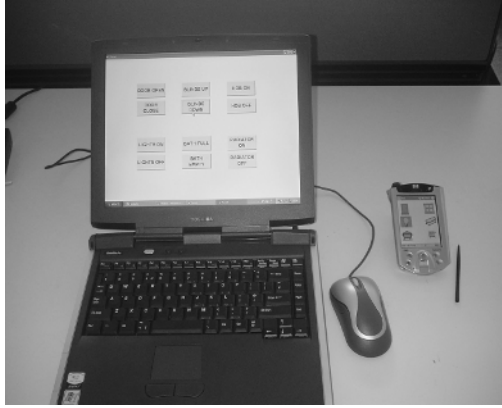
Our study compared the performance of participants when using three different modalities of output. These were: first a visual condition where the interface was comprised solely of icons, a second audio condition in which musical earcons were played in order to present an ambient style of interface, and a third condition in which a combination of icons and earcons were presented to the user, thus creating a multimodal interface. Participants tested the interface using two different apparatus: a laptop screen and a palmtop device.

## 6.1 Apparatus

Custom software was built for the study of the three experimental conditions. The software was implemented using Cakewalk along with Visual Basic for the laptop and embedded Visual C++ for the palmtop. The earcons were saved as MIDI files and the icons as JPG and GIF files. The experiments were carried out on a palmtop model HP iPAQ Pocket PC h5550 and a laptop model Toshiba Pro SP2100, mouse model Logitech M-UD43. See Fig. 5.

## 6.2 Subjects

The participants were comprised of 30 volunteers divided into two groups: a group of older users and a group of younger users. The group of older users was comprised of 6 males and 9 females with ages ranging between 60 and 84. The group of younger users covered ages ranging between 20 and 32 and was formed by 8 males and 7 females. None of the participants reported hearing problems, attention deficit problems or any other condition that would impede them in the performance of the task.



**Fig. 5.** Laptop with audio interface and palmtop used in the experiments



**Fig. 6.** Participant interacting with the palmtop device

## 6.3 Experimental Procedure

The experiment was carried on in 5 Phases. In the first phase, participants became familiar with the interfaces and apparatus and learnt how to interact with them. A maximum time of 20 minutes was allowed for them to learn and remember the earcons. In the second phase, participants were tested for response time and accuracy. In the third phase, which took place the next day, participants were tested again (retention). In the 4<sup>th</sup> phase, participants were introduced to two new representations

of two appliances and learnt them. In the last experimental phase, participants were tested again for time reaction and recognition. To avoid learning effects, both presentation order (modality) and device (apparatus) were randomly assigned to each participant. Similarly, the appliance chosen for interaction was also randomly assigned to the experiment so learning effects were not present.

The appliances only had two possible states; corresponding to on/off, open/close, up/down and full/empty, depending on whether they referred to the lights, door, blinds or bath respectively. For the fourth phase on the second day two new appliances, the radiator and the hob, were introduced.

During the visual portions of the exercise, participants were told to pay attention to the interface, since it would reflect any changes occurring in any of the appliances, and this could occur at any given time. It was explained that the information concerning the change would be reflected on the interface in the form of a change of icon and/or the emission of an earcon corresponding with the new status of the appliance, *i.e.* the icon representing the empty bath would be replaced by an icon representing the bath full when the change occurred. For the audio condition, the sound of the bath full would be played when the change occurred. In the experiment, participants were asked to identify the appliance and revert it to its original status by pushing the appropriate button. In a hypothetical situation, the subject might choose to ignore the message and leave the appliance in the new status. In the case of the audio only interface, the participant would know if he/she had pressed the correct button if the expected tune was played. Only when participants chose the correct button would changes occur. Identification of the right tune and/or icon was necessary to complete the task correctly.

#### 6.4 Data Collection

Participants took the test individually in front of the laptop screen that was resting on a table or given the palmtop to hold in their hands and proceed with the test, as shown in Fig. 6. The first trial in each modality was considered a training trial and, as such, was not taken into account when undertaking the statistical analysis. Both subsequent trials for each condition were recorded and averaged for the purpose of data analysis. Reaction time was measured by calculating the time elapsed between stimuli presentation (icon or earcon) and response (pressing the appropriate button).

Once participants had finished all tests, a semi-structured interview was conducted in which several questions about the devices and the modalities were asked as well as questions concerning safety and aesthetics.

#### 6.5 Results

Table 2 shows the mean, standard deviation, minimum and maximum times taken to respond to the different stimuli as presented in the two conditions either interacting with 4 or 6 representations of the appliances. From this descriptive data, it can be seen that the participants obtained slower performance times when presented with the ambient audio interface.

It can also be seen from Figure 7 that important differences were encountered during the analysis of the test results. The Wilcoxon test was used to evaluate the



results of the experiments. This statistical tests showed that there were significant differences in response times between the different modalities of output used during the test carried out with the palmtop using 4 different appliances. The results of the Wilcoxon test conducted show that the interface modality used had a significant effect on the time taken to react to changes. The mean of the response times for the audio only modality was slower than the visual ( $z = -4.286, p < 0.05$ ) and multimodal ( $z = -4.257, p < 0.05$ ) approaches. A significant effect was also found for the multimodal approach being faster than the visual mode only ( $z = -2.200, p < 0.05$ ).

Significant differences were encountered when comparing the mean times of the three output modalities displayed using the palmtop when interacting with 6 appliances. Participants performed faster when interacting with the visual ( $z = -2.172, p < 0.05$ ) and multimodal ( $z = -3.424, p < 0.05$ ) interface than when responding to the audio output. There was also a significant difference when comparing the mean of the multimodal and visual only modalities ( $z = -2.589, p < 0.05$ ).

**Table 2.** Experimental results in seconds

|            | Mean      | Std. Deviation | Minimum | Maximum |
|------------|-----------|----------------|---------|---------|
| Audio4Pal  | 22.211042 | 13.0515342     | 7.3650  | 54.1900 |
| Visual4Pal | 4.858750  | 1.3170531      | 3.7350  | 8.8550  |
| Multi4Pal  | 3.977708  | 4.1138935      | .4400   | 17.8100 |
| Audio4S    | 27.458575 | 13.9070187     | 13.3632 | 61.9727 |
| Visual4S   | 10.041813 | 9.7731297      | 3.7969  | 40.6269 |
| Multi4S    | 13.675682 | 8.7777750      | 5.5859  | 52.2910 |
| Audio6Pal  | 17.540952 | 19.7574405     | 4.9350  | 73.6400 |
| Visual6Pal | 8.261724  | 6.4471373      | 3.6100  | 24.7800 |
| Multi6Pal  | 6.598095  | 15.8792816     | .7200   | 75.2200 |
| Audio6S    | 27.332888 | 17.1831179     | 10.4296 | 69.0703 |
| Visual6S   | 7.470941  | 6.2761331      | 3.5664  | 29.6465 |
| Multi6S    | 11.671536 | 5.5968393      | 6.5156  | 27.5937 |

Equally significant were the results when analyzing the speed of response on the laptop display when interacting with 4 appliances. Participants performed faster using the visual ( $z = -4.453, p < 0.05$ ) and multimodal ( $z = -4.330, p < 0.05$ ) modes as compared with the audio only mode. In contrast with the palmtop, reaction times using the multimodal display were faster than with the visual modality. For the laptop, there were significant differences between the times taken to respond to the stimuli with the multimodal being slower than the visual mode ( $z = -2.684, p < 0.05$ ).

Similar results were obtained from analysis of the interaction with the 6-appliance interface on the laptop. The audio interface induced a much slower response rate than the visual ( $z=-4.623, p<0.05$ ) and multimodal modes ( $z=-4.554, p<0.05$ ). A significant difference ( $z=-3.416, p<0.05$ ) was also found between the visual and multimodal conditions, the multimodal being slower.

There were no significant differences when users were tested after having learnt 4 or 6 representations of the appliances.

It was found that there were significant differences ( $z=-3.743, p<0.05$ ) of performance among participants depending on the device used when testing the multimodal option and interacting with 4 appliances. Surprisingly enough, the screen modality was approximately three times slower ( $\mu= 13.6756$ ) than the palmtop ( $\mu=3.9777$ ).

Significant differences were also found between the performance of participants when using the palmtop and the laptop screen in the audio and multimodal systems when interacting with 6 appliances. A significant difference ( $z=-3.285, p<0.05$ ) was encountered in the time taken by participants using the multimodal condition, where the time taken using the screen ( $\mu= 11.6715$ ) was far greater than the time taken to respond to changes when using the palmtop ( $\mu=6.5980$ ).

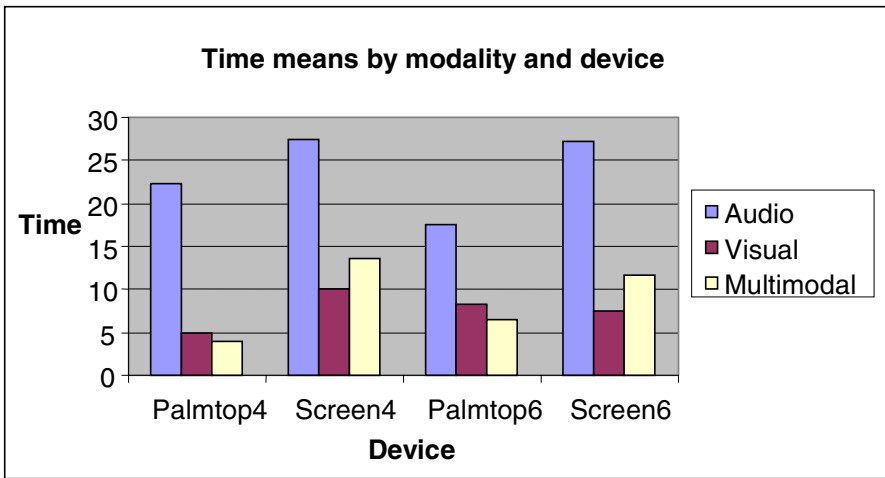


Fig. 7. Histogram of the means by modality and device

Also significant was the difference between the audio modes when displayed through the palmtop or laptop ( $z=-2.520, p<0.05$ ) test when interacting with 6 appliances' representations. In the aural mode, higher response times were recorded when interacting with the laptop ( $\mu =27.3328$ ) than when using the palmtop ( $\mu =17.5409$ ).

This is a surprising result, since it tells us in particular that users responded to an ambient interface more effectively when presented to them as a portable handheld form factor device. The audio aspects were effectively identical in the case of the palmtop as compared to the laptop, beyond superficial differences in the audio output

hardware on each device. We postulate that users were more comfortable using the palmtop as an ambient device due to its portable nature, associated more with mobile devices such as phones or remote controls, in contrast to the ‘fixed interface’ style of the laptop.

## 6.6 Interviews

The subjects’ positive comments about the laptop display were mainly related to the familiarity of it (5 people) and the larger area of display (8). Participants perceived it to be visually clear (6) and also recognized that the earcons were easy to look at on the screen.

Among the characteristics of the laptop display that were perceived as negative, participants mentioned the fact that it was fixed and thus could not be moved around as pleased them (7), but also that it was too big (3). The fact that the screen area was large was also perceived as negative by some participants (4) who reported a need to use peripheral vision to notice the changes in the icons.

Positive comments about the palmtop included its portability (15), easy of use (4) and convenience of use (2). Among the negative opinions were the need to use both hands to operate it (2), the fact that manipulating the buttons was harder than when using the laptop (2), the small size of the screen (4) and for new users (3) the need to practice.

## 7 Discussion

The first hypothesis has to be rejected in light of the results. In all conditions (palmtop and laptop screen, with 4 or 6 appliances) the responses of participants when using the audio interface were much slower than when interacting with either the visual or the multimodal interfaces. Therefore, we can safely say that the audio interface was not as efficient as the visual or multimodal interfaces in terms of response time.

These results are complementary to the opinions expressed by the users in the interviews after the experiments. In these interviews participants revealed the intense effort (in some cases) they went through during the learning process for the earcons. In contrast, the learning of the icons was immediate, with only a couple of participants needing clarification on the icons representing the radiator, all other icons were self-explanatory and needed no clarification from the experimenter.

This problem with the learning curve has already been established by Czaja *et al.* [65], but in a natural environment where the device will be used on a regular basis and users can interact with it at leisure it is possible that users will learn the earcons in a comparatively short period of time. Further research is needed in this area, and we cannot rule out additional benefits that stem from the use of audio as an ambient technique over visual alone.

The second hypothesis was accepted. The fact that participants did not improve or worsen notably during the second phase of the experiment meant that learning did occur during the experimental procedure. We can infer this from the fact that 2 new representations of appliances were introduced during the second day and performance did not subsequently worsen.

The third hypothesis was also accepted. The fact that there was no significant difference in performance between the group of elders and the group of younger users suggests that the design of the earcons and icons did not offer disadvantages to the group of elderly users. We can attribute this to the various techniques used during the design process of the interface and particularly to the inclusive design techniques that assured the appropriateness of the design for the group of elders.

In the case of the fourth hypothesis, we have to reject the null hypothesis as there were significant differences between the use of the palmtop device and the laptop screen in the speed of response. The palmtop device was consistently faster in all those conditions where audio and multimodal presentations were compared except the comparison between the audio modality with only 4 appliances. There were no significant differences in the visual only display. This ruled out any possible differences due to the device in itself. We therefore attribute these differences to the users' perception of the device as being more suitable for an ambient style of interface, where interpreting audio output is expected.

A further possible explanation of such differences could come from the difficulty participants encountered with the retention of earcons during the first day. More research will be needed to clarify the effects of device on performance.

## 8 Conclusions and Further Work

The results of the experiments showed that the multimodal interface was an appropriate way of delivering the requested information to the user. Taking into account the advantages offered by the multimodal approach over the visual modality, we would recommend prioritizing the former choice. In a real setting, we think elderly users would benefit in their daily living from a device that uses a multimodal interface to control the appliances in their home. Whilst users found it difficult to remember the sounds associated with each device, the results show that it does nonetheless aid recognition and response times. In addition, the ambient nature of the audio element provides benefits beyond an aid to recognition, such as the benefit gained from avoiding the need for constant concentration on the interface in order to observe status changes. The home can be made safer for residents if a multimodal household appliance controller is installed.

Taking into account the results of the experiment, with the implications for efficiency and previous knowledge from the literature, it can be said that to achieve a good degree of efficiency and acceptability it is essential that the user feels comfortable with the appliance. From the comments made after the test, it was clear that the portability of the palmtop offered a great advantage to the elderly users. It offered fast access and being portable increased the sense of safety for the elderly participants. It also offered the best performance in conjunction with the multimodal condition, which made this option the most efficient.

Using very clear icons that were easy to remember was essential for the good performance of the visual display. Similarly, the introduction of well-known tunes that were relatively easy to remember improved user performance. The redesign of the audio interface by introducing tunes that were familiar to the users, and therefore required less cognitive effort to learn and remember, improved drastically the rates of

correct identification. Providing interfaces that are similar, or resemble previous apparatus, will ease the learning process for operation of the system. Despite all efforts to make them easier to remember and recall, the earcons were not as efficient as the icons. However, due to the fact that participants were limited in the time they had to learn the earcons and the fact that in a real situation the learning process would take place in their own time and conditions, it is not hard to imagine that the audio only mode could eventually become an efficient alternative. Certainly for users with visual impairments it would constitute a good substitute. Nevertheless, the results confirm the belief that a multimodal interface better serves the purpose of adaptation to new personal situations, whether it be physical decline, activity engagement by the user, *etc.* Further research into the possibilities of introducing a different audio display is necessary.

## References

- [1] Edwards, W.K., Grinter, R.E.: At home with ubiquitous computing: seven challenges. In: G.D. Abowd, B.B., S.A.N.Shafer (ed.): Ubicomp. Springer-Verlag, Atlanta, USA (2001) 256-272
- [2] Jacobson, D.R.: Representing Spatial Information through Multimodal Interfaces. Sixth International Conference on Information Visualisation (IV'02). IEEE, London (2002) 730-736
- [3] Reeves, L.M.: Guidelines for multimodal user interface design. Communications of the ACM 47 (2004) 57-59
- [4] Alm, N., Arnott, J.L., Dobinson, L., Massie, P., Hewines, I.: Cognitive prostheses for elderly people. (2001) 806-810
- [5] Hawthorn, D.: Possible implications of aging for interface designers. Interacting with Computers 12 (2000) 507-528
- [6] Edwards, W.K.: Enabling technology for users with special needs. IHM-HCI, Vol. Tutorial 10, Lille, France (2001) 1-83
- [7] Arnold, M., Hopewell, P., Parry, P., Sustache, N., Paddison, C.: User Centred Design - How to Design Accesible Products. In: Maguire, M.A., K. (ed.): European Usability Professionals Association Conference., Vol. 3. British Computer Society, London (2002) 22-31
- [8] Statistics, N.: Social and welfare., Vol. 2002. National Statistics (2002)
- [9] Kalache, A., Lunenfeld, B.: Health and the ageing male. World Health Organisation, Geneva. (2001)
- [10] Tinker, A.: Older people in modern society. Longman, London (1997)
- [11] Jorge, J.: Adaptive Tools for the Elderly New Devices to cope with Age-Induced Cognitive Disabilities., Vol. 2003. WUAUC01 (2001)
- [12] Whitney, G.: Sensory augmentation system. The use of technology to provide complementary information for people who have a sensory impairment to enable them to travel safely and comfortably., Vol. 2002. UK Computing Research Committee. (2002)
- [13] Fozard, J.L., Gordon-Salant, S., Scheiber, F., Weiffenbach, J.M.: Sensory and perceptual considerations in designing environments for the elderly. In: AARP (ed.): Life-Span Design for Residential Environments for an Aging Population., Vol. 2001. HMRC, Washington D.C. (1993)
- [14] INCLUDE: INCLUSION of Disabled and Elderly people in telematics. Vol. 2002 (2002)

- [15] Zhao, Y., Tyugu, E.: Towards a personalized browser for elderly users. In: Waern., S. (ed.): Workshop on user interfaces for all; towards an accesible Web., Långholmen, Stokolhm (1998)
- [16] Motluk, A.: Infinite sensation. *New Scientist* (2001) 24-28
- [17] Shiffman, H.R.: Sensation and perception. An integrated approach. John Wiley & Sons, INC., New York. (2001)
- [18] Freeland, A.P.: Deafness: the facts. Oxford U.P., Oxford (1989)
- [19] Lysons, K.: Understanding hearing loss, Jessica Kingsley (1996)
- [20] Stevens, J.C., Cruz, L.A.: Spatial acuity of touch: ubiquitous decline with aging revealed by repeated threshold testing. *Somatosensory & Motor Research* 13. (1996) 1-10
- [21] Sathian, K., Zangaladze, A., Green, J., Vitek, J.L., DeLong, M.R.: Tactile spatial acuity and roughness discrimination: impairments due to aging and Parkinson's disease. *Neurology* 49 (1997) 168-177,
- [22] Sharps, M.J.: Age-related change in visual information processing: toward a unified theory of aging and visual memory. *Current Psychology: Developmental Learning Personality Social* 16 (1998) 284-307
- [23] Hawthorn, D.: Cognitive and human computer interface design. In: IEEE (ed.): Australasian Computer Human Interaction Conference. IEEE, Adelaide, S. Australia (1998) 270-281
- [24] Brewster, S.A., Wright, P.C., Edwards, A.D.N.: Experimentally derived guidelines for the creation of earcons.: HCF'95, Huddersfield (1995)
- [25] Vickers, P., Alty, J.L.: Towards some organising principles for musical program auralisations. In: ICAD'98 (ed.): ICAD'98. ICAD'98, Glasgow (1998)
- [26] Dong, H., Keates, S., Clarkson, P.J.: Accomodating older users' functional capabilities. In: Brewster, S., Zajicek, M. (ed.): HCI BCS 2002, London (2002) 10-11
- [27] Dewsbury, G., Taylor, B., Edge, M.: Designing safe smart home systems for vulnerable people. In: Rouncefield, R.P.a.M. (ed.): Dependability in Healthcare Informatics. Lancaster University, Edinburg (2001) 65-70
- [28] Hanson, V.L.: Web AAccess for elderly citizens. WUAUC'01. ACM, Alcacer do Sal, Portugal (2001) 14-18
- [29] Dewsbury, G.: The social and psychological aspects of smart home technology within the care sector. *New Technology in Human-services* 14 (2001) 9-17
- [30] Mynatt, E.D., Essa, I., Rogers, W.: Increasing the opportunities for aging in place. *ACM Proceedings of the 2000 Conference on Universal Usability* (2000) 65-71
- [31] Baldock, J., Hadlow, J.: Housebound Older people: the links between identity, self-esteem and the use of care services. ESRC. Growing Older Programme., Sheffield (2002) 1-4
- [32] Peace, S., Holland, C., Kellaher, L.: Environment and Identity in later Life: a cross-setting study. ESRC. Growing Older Programme., Sheffield. (2003) 1-4
- [33] Regnier, V.: Design principles and research issues in housing for the elderly. In: AARP (ed.): Life-Span Design for Residential Environments for an Aging Population., Vol. 2001. HMRC, Washington D.C. (1993)
- [34] Allen, B., Ekberg, J., Willems, C.: Smart houses: how can they help people with disabilities? In: R.W.Roe, P. (ed.): telecommunications for all. ECSC-EC-EAEC, Brussels (1995)
- [35] Chan, M., Hariton, C., Ringear, P., E., C.: Smart house Automation system for the elderly and the disabled. (1995)

- [36] Monk, A.F., Baxter, G.: Would you trust a computer to run your home? Dependability issues in smart homes for older adults. In: Brewster, S., Zajicek, M. (ed.): BCS HCI 2002, Vol. A new research agenda for older adults, London (2002) 21-22
- [37] Sainz Salces, F.J.: Multimodal human-computer interaction. HCI 2002, South Bank University, London (2002)
- [38] Barrass, S., Kramer, G.: Using sonification. *Multimedia Systems* 7 (1999) pp 23-31
- [39] Brewster, S.: Providing a structured method for integrating non-speech audio into human-computer interfaces., Vol. 2000 (1994)
- [40] Arroyo, E., Selker, T., Stouffs, A.: Interruption as multimodal outputs: which are the less disruptive? : International Conference on Multimodal Interfaces. IEEE (2002)
- [41] Fraser, J., Gutwin, C.: The Effects of Feedback on Targeting Performance in Visually Stressed Conditions. *Graphics Interface '00.*, Vol. *Graphics Interface* (2000) 19-26
- [42] Hooegeven, M.: Towards a new multimedia paradigm: is multimedia assisted instruction really effective? , Vol. 2001 (1995)
- [43] Makris, P.: Accessibility of Ubiquitous Computing: Providing for the Elderly. Vol. 2003. Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly (2001)
- [44] Savitch, N., Freeman, E., Clarke, L., Zaphiris, P.: Learning from people with dementia to improve accessibility of website interfaces. In: Dearden, A., Watts, L. (eds.): HCI2004, Vol. 2. British HCI Group, Leeds (2004) 185-186
- [45] Wilson, C.M., Lodha, S.K.: Listen : a data sonification toolkit. Vol. 2001. ICAD'96 (1996)
- [46] Ng-A-Tham: Equality service accessible for all citizens, in particular elderly and disabled.: TIDE (1998)
- [47] Richter, K., Enge, M.: Multimodal framework to support users with special needs in Interaction with public information systems. In: L.Chittaro (ed.): The human computer interaction handbook. Oviatt, S. (2003) 286-301
- [48] Venkatesh, A.: Computers and other interactive technologies for the home.: *Communications of the ACM*, Vol. 39 (1996) 47-54
- [49] Bien, Z.Z., Park, K., Bang, W., Stefanov, D.H.: An intelligent sweet home for assisting the elderly and the handicapped. CWUAAT (2002)
- [50] Ifukube, T.: A neuroscience-based design of intelligent tools for the elderly and disabled. WUAUC'01. ACM, Alcaccer do Sal (2001)
- [51] Blackwell, A.F., Hague, R.: AutoHAN: an architecture for programing the home. IEEE Symposia on Human-Centric Computing Languages and Environments (2001) 150-157
- [52] Park, S.H., Won, S.H., Lee, J.B., Kim, S.W.: Smart home – digitally engineered domestic life. *Personal Ubiquitous Computing* 7 (2003) 189–196
- [53] Wan, D.: Magig Medicine cabinet:a situated portal fo rconsumer healthcare. First International Symposium on Handheld and Ubiquitous Computing (HUC '99), Karlsruhe, Germany (1999)
- [54] Mann, W.C., Ottenbacher, K.J., Fraas, L., Tomita, M., Granger, C.V.: Effectiveness of assistive technology and environmental interventions in maintaining independence and reducing home care cost for the frail elderly. *Arch Fam Med* 8 (1999) 210-217
- [55] Bonner, S.: Assisted Interactive Dwelling house. Vol. 2003. TIDE (2003)
- [56] Linden, v.d.: Present Practices. Vol. 2003. European Union (2001)
- [57] van Berlo, A.: A "smart" house as research and demonstration tool for telematics development. Vol. 2002. STAKES (1998)

- [58] Chan, M., Bocquet, H., Campo E., Pous J.: Remote Monitoring System to Measure Indoors Mobility and Transfer of the Elderly. *Technology for Inclusive Design and Equality.TIDE.*, Vol. 2002, Helsinki (1998)
- [59] Sainz Salces, F.J., England, D., Llewellyn-Jones, D.: Designing Ambient Home Interfaces for Elderly People. In: B. Sierra, E.L. (ed.): *Workshop on Ambient Intelligence and (Everyday) Life*. Springer, San Sebastian, Spain (2005) 205-214
- [60] Kohler, M.: A vision based hand gesture recognition system for controlling appliances in the intelligent house., Vol. 2002
- [61] Vallés, M., Manso, F., T., A.M., Del Pozo, F.: Multimodal environmental control system for elderly and disabled people. In: IEEE (ed.): *18th Annual Conference of the IEEE Engineering in Medicine and Biology Society.*, Amsterdam (1996)
- [62] Chan, M., Bocquet, H., E., C., J., P.: Remote Monitoring System to Measure Indoors Mobility and Transfer of the Elderly. *Technology for Inclusive Design and Equality.TIDE.*, Helsinki (1998)
- [63] Bühler, C., Clemens, D., Heck, H., Wallbruch, R.: The KommAS Communication Aid for the Elderly people with aphasia.: TIDE, Helsinki, Finland (1998)
- [64] CNN: Panasonic banks on digital for the elderly., Vol. 11/03/02. CNN (2002)
- [65] Czaja, S., Clark, C., Weber, R., Nachbar, D.: Computer communications among older adults. In: society, H.f.a.e. (ed.): *Proceedings of the Human factors and ergonomics society, 34th Annual Meeting.* (1990) 146-148
- [66] Brownsell, S., Williams, G., Bradley, D.A.: Information strategies in achieving an integrated home care environment. In: IEEE (ed.): *Serving Humanity, Advancing technology.*, Atlanta (1999) 1224
- [67] Shao, J., Tazine, N., Lamel, L., Prouts, B., Shröter, S.: An open system architecture for a multimedia and multimodal user interface., Vol. 2002
- [68] Lines, L.: Designing spoken dialogue for intelligent home system. In: J. Vanderdonckt, A.B., A. Derycke (ed.): *IHM-HCI 2001, Interaction without frontiers- Interaction sans frontières.*, Vol. II. Cèpaduès-Editions, Lille, France. (2001) 187-188
- [69] Zajicek, M., Morrisey, W.: Spoken message length for older adults. INTERACT. The Speech Project, Tokyo, Japan (2001)
- [70] Morrissey, W., Zajicek, M.: Can sound output enhance graphical computer interfaces? In: Hanson, M. (ed.): *Contemporary Ergonomics*. Taylor and Francis, London (2000)
- [71] Jedamzik, M.: Smart House. A usable dialog system for the control of technical system by gesture recognition in home environments. *Technical Possibilities, State of the Art in Technique and Research.*, Vol. 2003 (1995)
- [72] Keates, S., Clarkson, P.J., Robinson, P.: Developing a practical inclusive interface design approach. *Interacting with computers* 14 (2002) 271-299
- [73] Sainz Salces, F.J., England, D., Vickers, P.: Household appliances control device for the elderly. In: Brazil, E., Shinn-Cunningham, B. (eds.): *ICAD*. Boston University Publications, Boston, USA (2003) 224-227



# A Smart Electric Wheelchair Using UPnP

D. Cascado, S. Vicente, J.L. Sevillano, C. Amaya,  
A. Linares, G. Jiménez, and A. Civit-Balcells

ETS Ingeniería Informática. Universidad de Sevilla.  
Av. Reina Mercedes, s/n. 41012, Sevilla, Spain  
danic@atc.us.es

**Abstract.** People with disabilities in general, and wheelchair users in particular, are one of the groups of people that may benefit more from Ambient Intelligent (AmI) Systems, enhancing their autonomy and quality of life. However, current wheelchairs are usually not equipped with devices capable of accessing services in AmI environments. In this paper, we describe how an electric wheelchair is equipped with an UPnP based module that allows the integration in AmI systems.

## 1 Introduction

Although the Ambient Intelligent (AmI) concept is not oriented towards any particular group of people, it is obvious that the AmI emphasis on greater user-friendliness, more efficient services support, user-empowerment, and support for human interactions [1] would be especially useful for people with disabilities and elderly people. In this paper, we focus on the integration of wheelchair users in AmI systems. Consider, for instance, the following scenario:

*A wheelchair user with several mobility restrictions and a mobile computer in his/her wheelchair enters a building (let's say the rehabilitation centre), provided with Ambient Intelligent facilities. As soon as he/she gets into the building, the Ambient Intelligent System (AmIS) discovers his/her presence and the devices announce the services that can be used, according to his/her special needs (cognitive, sensorial, physical and communication abilities), and technological constraints (display resolution, voice, text, pixel-based, bandwidth, computing power, etc.).*

*The AmIS offers communication with a remote information centre that appears in the user's display, adapted to his/her physical and cognitive characteristics (text menu, voice, icons...). The AmIS offers information about where he/she is and where to go from the current position and, after knowing where to go, the possibility of a route-guiding tool appears on its display, which the user accepts. It uses a location service and gives information via text messages. The user is located and receives his/her position, together with a message of where to go now. The AmIS has calculated the best path to follow, taking into account the user constraints (for instance, avoiding stairs, changing the timing of automatic doors, etc.) and the information of occupation in the building at that time of day.*

Finally he/she arrives at the destination and the AmIS sends information about what domotic devices are installed in the room (and can be used). From its joystick with buttons (or any other adapted input device), the user sends orders to control devices offered by the domotic system (i.e., switch the lights off or on, roll up or down blinds, change the television channel, set the temperature of the air conditioning system, etc.). Afterwards, he/she decides to leave the building, but this time the route-guiding tool is not employed, and the joystick is used to guide the wheelchair to the exit.

Note that some characteristics of Ambient Intelligent systems are particularly well suited for people with mobility restrictions:

- Ubiquitous access: allows access to services in a way that is not restricted by the location of resources and/or the user's mobility (remote control for TV, air conditioning, etc.; answering the phone from the wheelchair, etc.).
- Context awareness: apart from the obvious use of location awareness, other dimensions are useful in our case, particularly personal awareness (dynamic adaptation to user needs, abilities or preferences) [2].
- "Invisible" computing and networking: allows un-noticed user monitoring in terms of safety: falls, care for people who may get lost, etc. (e.g. elderly residences).

However, since many of these people usually require the use of mobility aids (like wheelchairs) and adapted user's interfaces, these assistive devices should also be integrated into the AmI system. For instance, it may be useful for the handicapped to access assistive services through their personalized interfaces (especially in unfamiliar environments). Furthermore, wheelchairs should be able to use and provide services (location, semiautomatic navigation, etc.). However, current wheelchairs are usually not equipped with devices capable of accessing services in AmI environments. At most, a wheelchair user may carry a portable computer or a PDA to access services, but in this case, the wheelchair itself is not integrated in the AmI system.

In this paper, we describe an electric wheelchair that is equipped with a module that allows the integration in AmI systems. We first describe a hardware module that serves as an interface between the wheelchair and the external devices. In the next sections, we explain the software architecture, which is based on UPnP<sup>1</sup> (Universal Plug and Play), and we focus on the User Interface software. Finally, we present the conclusions.

## 2 Hardware Architecture

The wheelchair used in this work is based on a former prototype called *Tetranauta* [3,4,5], a low cost, fully open steering system that allows people with severe motor impairments to move in known environments: hospitals, schools, home, etc. Navigation is assisted by allowing the wheelchair to follow predefined paths (with tracks

---

<sup>1</sup> <http://www.upnp.org>

marked on the floor) and also by using an infrared-based obstacle detection system. As a result, the user effort and safety in driving the wheelchair, especially in long paths, is improved.

So far, the majority of the efforts in wheelchairs have been oriented to empowering the autonomous capabilities, like obstacle avoidance, trajectory tracking, efficient suspension of wheels or stair-climbing capabilities. An example of this is OMNI [31] that included obstacle avoidance, human-machine interface, high maneuverability and navigational intelligence. NavChair [25] is a smart wheelchair that is able to avoid obstacles, and to follow a direction indicated by a user with tremor or another type of severe mobility impairment. IBOT-3000 [29] was capable of climbing stairs with its two pairs of balanceable wheels. The Smart wheelchair of the CALL Centre [26] allows several types of interaction modes with the user, according to his/her disability and skills. The wheelchair Tetranauta was in this line of developments, being an isolated wheelchair without communications with another devices. In [18], a revision of this type of wheelchairs can be found.

The next generation of wheelchairs contemplates smart capabilities for autonomous operation and communication capabilities with other systems (like wheelchairs, domotic devices or the like). In this sense, electric wheelchairs have been used like test beds for communication systems, like E-wheelchair [27] that was used to prove the viability of IPv6 communications. The design shown in [28] used its wireless communications to improve its autonomous capabilities (to communicate with a GPS to obtain the position of the wheelchair).

However, these last ones were an attempt of expanding the capabilities of wheelchairs through communications, but not to integrate the wheelchair in a bigger system. In this line, AmIChair [30] can use communications to control the devices of the environment, but environment devices can also monitor and control the wheelchair. The wheelchair is integrated in the whole system and it is only a part of it. Our development pursues a similar idea: being a device integrated in a bigger system, but now the user can interact with the system using its proper wheelchair's adapted interface.

The control unit of the wheelchair is composed of an embedded computer, plus several functional modules controlling different parts of the wheelchair: power module, steering device module, etc. All these modules are inter-connected by a DX serial bus [6], a *de facto* standard in electric wheelchairs. As a result, the wheelchair becomes a distributed embedded system where new functional modules may be connected with relatively few software changes. Particularly, in this paper we describe how a new gateway/bridge module is incorporated into the wheelchair so that the DX sub-systems can communicate with other devices in an AmI environment.

The system architecture is shown in Fig. 2. Essential DX modules are the User Control Module (UCM) and the Power Module (PM). The UCM is normally part of a control module (which includes a joystick or any other type of speed and direction control), and has the function of processing signals from the wheelchair user and sending instructions to the other modules. These instructions and other data are sent using messages that are named Network Variables (NV). The PM provides the controlled voltage to drive the wheelchair's motor(s) and operates the park brakes. These are modules already included in any DX-based wheelchair.

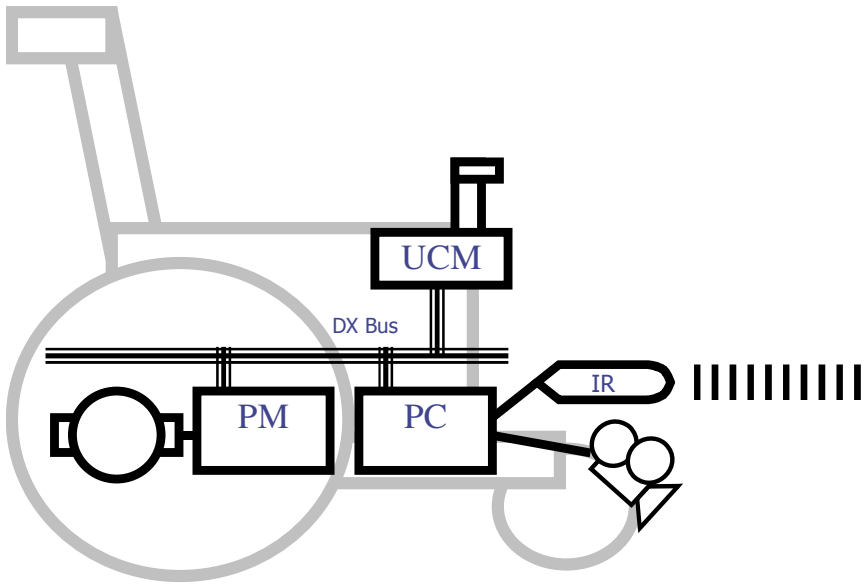


Fig. 1. Basic prototype's scheme

From our point of view, the central element is the user interface (UI), not included in standard DX wheelchairs. The UI should have the following characteristics:

- It should be a mobile system, easily handled by the user (who may not always be seated in the wheelchair) or any other person (relative, carer, nurse, etc.).
- It should have enough resources (computing power, memory, screen size, bandwidth) to run useful (visual) applications adapted to the user's needs and preferences. Furthermore, development kits should be available in order to write application-specific software.
- It should incorporate multiple communication links, preferably wireless, to allow ubiquitous access to other devices and services.

Advancements in handheld devices (such as PDAs, mobile phones and portable PCs) are enormous and there are now commercially available devices of acceptable performance at relatively low cost. In our prototype, we decided to use a StrongARM® based PDA as our user interface (UI), running under WindowsCE® 3.0. The UI is equipped with a color screen, as well as with serial, Bluetooth and 802.11 interfaces. Although usually a wireless connection is more adequate, sometimes a simple serial (wired) link is a simpler and more robust solution (for instance, when the PDA is attached to the wheelchair).

Another added element is the *DX-Bridge*, a new DX module that captures the Network Variables (NVs) flowing through the DX bus. This module allows NVs (for instance, wheelchair control data) to be exported as input for the user interface (UI), as well as to receive new values for the NVs from the UI. So far, the only allowed way to capture DX data from the bus is using a device named DX-KEY, provided by

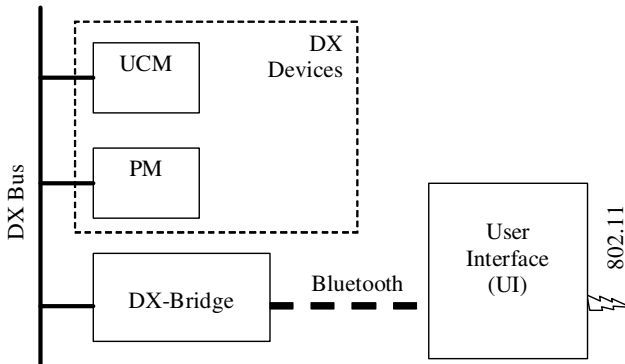


Fig. 2. Scheme of wheelchair’s hardware

Dynamic Controls [7]. This is a DX module that provides an interface between the DX bus and an external system, allowing access through a parallel port to DX variables. Therefore, the DX-Bridge is a hardware module based on a Cygnal C8051F330 microcontroller [8] with serial and parallel interfaces: the DX-KEY is accessed through the parallel port, and on the other side a serial link (RS-232, 115200bps) is used to connect with the UI. As we said before, some times this wired link is enough. However, in order to provide a wireless link, we also use a commercial serial-Bluetooth module that works under the RFCOMM profile (Serial Cable Emulation).

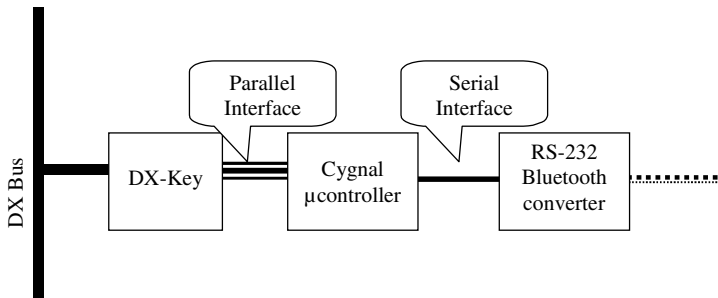


Fig. 3. Hardware in the DX-Bridge

The DX-Bridge has two operation modes: *driving*, and *domotic* modes. In the driving mode, the DX-Bridge simply captures the commands (NVs) sent by the User Control Module (UCM) and then it delivers these commands to the Power Module (PM). These commands are delivered without any changes, so the wheelchair operation is not different from a standard manually operated wheelchair. In the domotic mode, the DX-Bridge captures the commands sent by the UCM and then it delivers these commands to the user interface-UI. In this way, these commands are no longer used to control the wheelchair, but they are interpreted by the UI as commands to control a domotic system or the like. Obviously, the domotic mode cannot be used unless the wheelchair has reached a secure state, avoiding dangerous situations for the user.

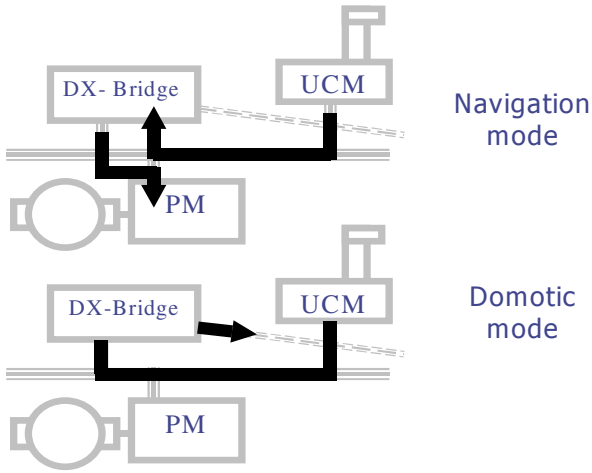


Fig. 4. Operation modes of DX-Bridge

The use of the wheelchair UCM to control a domotic system has two main advantages: first, the user is probably used to handling the control device (e.g. joystick) to drive his/her wheelchair, and therefore he/she would probably learn to use the external devices more easily. And second, since the UCM would probably be adapted to the user’s needs and/or difficulties, we have an adapted control device for domotic systems “for free”. This is a key question because with this solution, the cost of domotic systems does not depend on the user’s physical and/or cognitive abilities.

Finally, although it cannot be considered as an operation mode, the DX-Bridge may be used by the UI to set new values for some DX variables. For instance, if an external service provides positioning and location information, these data can be sent to an optional assisted navigation DX module.

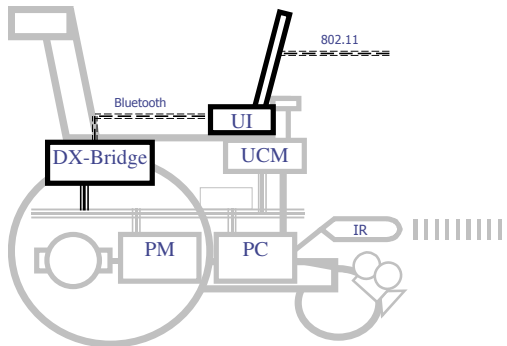


Fig. 5. Final architecture of the wheelchair

### 3 UPnP Architecture

As we said before, the wheelchair must be capable of obtaining information from the environment, for instance positioning information or the list of available devices and/or services, and the domotic system (or the like) must be capable of obtaining information from the wheelchair (like orders to the different devices of the room). Connection between these devices can be implemented in several ways: infrared cards, radio communication systems, or even wired connections. In our system, communications are centered on the user interface (UI), which should be able to communicate across heterogeneous and dynamically changing links and networks. For instance, the UI would be able to communicate with data networks, Internet Access Points, domotic buses like EHS (a Powerline-like bus used for the control of home devices) [9], etc. Wireless personal (e.g. Bluetooth) and local (e.g. Wi-Fi) area networks now permit low-cost commercial solutions for this type of communication, but there are still open problems like efficient roaming, reachability, intermittent failures, fault tolerance, security, etc.

Note that it is not only a problem of *interconnectivity* or interactions at the inter-networking level, but also of interaction among devices at higher levels: control, configuration and information sharing in different formats, import/export services, etc. Among the different communication architectures available (Juni, UPnP, HAVi), as discussed in [10], we consider that UPnP is a good choice for implementing our communication system. UPnP is a lightweight set of protocols to extend the Plug & Play concepts to network devices, and it supports all mentioned functions including the dynamic connection of a device to a network, services offering and discovery, everything based on a unified description of functions and attributes of services through XML (*eXtended Mark-up Language*) documents [11]. UPnP is capable of working with scarce resources and unreliable connections (devices can suddenly appear and disappear), and a further reason to choose UPnP is that there are a large number of available SDKs for several platforms and operating systems [12,13]. Furthermore, there are two factors that make UPnP especially attractive from our point of view: one is the use of open and standard protocols; second is the use of the IP protocol at the lowest level.

Indeed, IP protocol has demonstrated its success in the interconnection of heterogeneous devices (a good example is the Internet). Most devices can be connected through a backbone IP network while secondary, maybe simpler, devices (e.g. sensors) may be connected using non-IP communications. In this case, a gateway is used to interconnect IP and non-IP sub-networks. For instance, in our prototype, we need gateways to interconnect the IP backbone network to the DX bus (a non-IP control network). Furthermore, UPnP operates with a set of existing and well-tested protocols and only needs an auto-IP network for running. UPnP works in a distributed philosophy, and classifies devices into two roles: *control points* or clients, and *host devices* or servers of services. However, almost all UPnP devices implement both client and sever functionalities, so peer to peer communications are possible.

For the wheelchair to be integrated into an AmI environment, the wheelchair must export information, acting as a host device: we need to know the position of the joystick, if any button of the console was pressed, the battery status, and so on. On the other hand, the wheelchair also needs to act as a control point (client): it has to know

its location in order to know how to get to another room/place, selecting the route from a map, etc. All these information/services are supplied from other devices (services) in the network. Since not only the wheelchair, but also the domotic system and most other elements of the AmI system get information from other devices, control points must be implemented on them.

The role of the user interface (UI) is to serve as an interface between the wheelchair and these UPnP services. All devices implementing UPnP services (host devices) should be connected by means of a backbone IP-based network. In our system, since the wheelchair is a mobile system, and services should be accessed "on the move", we use an IP wireless network (802.11b/g [14]).

Among the many different host devices that may be present in an AmI system, we identify the following for our scenario (see Fig. 6):

- **Monitoring:** exports wheelchair's status variables (i.e.: joystick position) and it is implemented in the wheelchair's gateway through serial (or Bluetooth) interface and the DX Bridge.
- **Map storing:** a small processor with an associated memory for storing and reading maps. This device can be in a fixed place of the building.
- **Location:** offers a positioning service, maybe out-doors positioning (like a GPS device attached to the wheelchair), or in-doors positioning [15].
- **Domotic:** offers as a service the kind of operations that may be performed with the associated domotic devices, tells which are the available devices at a specific room. An UPnP control point attached at the same domotic service's host device can read the status of the wheelchair to generate an alarm in the domotic system if the battery is low or can read the position of the joystick for using it as an input device to handle a graphical interface of any domotic device.

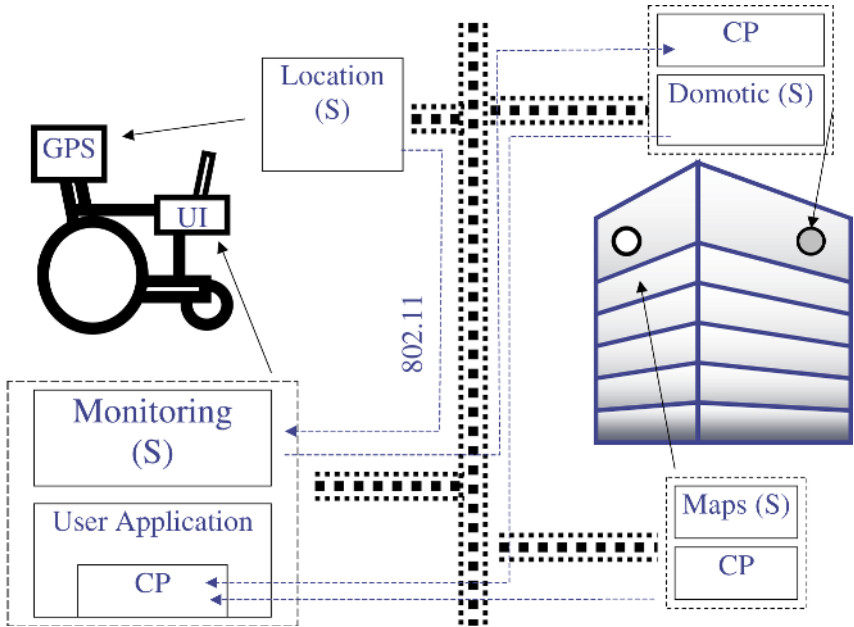
## 4 The User Interface (UI)

In this section, we describe the implementation of the UPnP host device in the user interface (UI). This description serves as an example, since all host devices can be implemented in a similar way. The UI has been implemented over a StrongARM®-based PDA equipped with serial, Bluetooth and 802.11b interfaces. The latter is used for supporting UPnP activity, as described in the previous section. We use an under-request protocol implemented *ad hoc*, that works as follows. The UI periodically sends a set of inquiry frames to the DX sub-system to know what the values of the DX variables are. When the micro-controller in the DX-Bridge (see section 2) receives an inquiry frame, it returns the value of the inquired variable. We preferred this simple solution because the Cygnal micro-controller used in the DX-Bridge is not powerful enough to support the UPnP stack. However, we could have implemented an UPnP DX variable service at an increased cost and complexity.

The UPnP software in the UI runs as an application over WindowsCE® 3.0. Basically, the application contains two protocol stacks: Serial communication stack and UPnP stack (see Fig. 7). The former implements the communications protocol between the DX-Bridge and the UI. This protocol is very simple; it only needs PHY and MAC layers, including some error correction capabilities like frame retransmission.



Over these two layers, a layer named *DX variable store* is implemented in order to guarantee the consistent storing of values (note that these DX variables are "resources" accessed both by the Serial Communication Stack and by the higher layer, so this layer has to cope with the consistent use of these shared resources). On top of both stacks is the UPnP service layer, which implements two threads: a DX inquiry loop (that gathers DX variables from the DX variable store layer) and the UPnP processing thread (responsible of gathering all the UPnP requests and generating the appropriate responses). DX inquiry loop notifies when a DX variable changes in order to notify this change to the UPnP clients connected to the service.



**Fig. 6.** UPnP services (S), Control Points (CP) for UPnP architecture. Dots indicate data flows.

The main task of UPnP stack is the implementation of the UPnP service. This is composed of a set of actions (methods) and status variables that the UPnP client (control point) can invoke at any time.

Service's status variables are all the DX variables that are desired to be monitored: joystick position; button, battery, DX control unit and serial communication status, and so on. The UPnP Monitoring Service maintains an inquiry loop responsible for holding the latest values of these variables. On the other hand, these services' status variables can be read by other UPnP devices in the IP network under demand (this is the default mode) or by a change notification event. Under this latter mode, the client receives an event every time the DX variable changes. When this event occurs, the client receives the variable name and value.

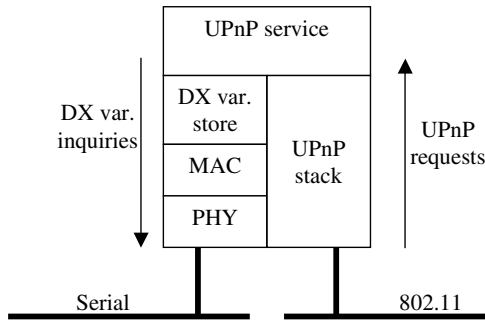


Fig. 7. UPNP application architecture in wheelchair's gateway device

Additionally, actions included in the service allow to read/write all the DX variables (not only the service's status variables) and even some internal DX variables that can be identified by its DX identification number (note that all DX variables have an identification number). However, these hidden internal DX variables cannot be read by events. Additionally, there's an action that resets the DX stack and the micro-controller (like a warm-reset).

Finally, the UPNP layer also implements a web interface, which allows all the status variables to be read and all the actions described above to be executed by means of a web page (see Fig. 11). As a result, the system is not only accessible as an UPNP device, but it could also be accessed as a simple web-controlled device. Since the wireless backbone IP network may be connected itself to the Internet with a *Residential Gateway*, remote control or maintenance via a web page is allowed.

## 5 Final Remarks

Some aspects have to be taken into account when using the wheelchair as described above. First, remember that commands sent by the UCM (i.e., the wheelchair's joystick) can be captured by the DX-Bridge and delivered to the user interface-UI in what we called the *domotic* mode. These commands are no longer used to control the wheelchair, but they are interpreted by the UI as commands to control an external system. Since in these cases the wheelchair acts as an UPNP host device (*monitoring service*), all these commands are transmitted through the UPNP protocol stack, including IP. Due to the high response times of these protocols, only very simple interfaces can employ the monitoring service as an input device. For instance, if we want to use the joystick movements to control an external domotic system, the corresponding changes of DX variables should be sent through UPNP. Tracking the joystick movements to control something similar to a mouse cursor would not be possible since it requires to send an event every time the DX variable changes.

However, DX-variables containing joystick positions can be used directly for controlling the user interface, since communications between the DX Bridge and the UI are performed through a serial/Bluetooth link and not through UPNP. For instance, in

our prototype, the joystick can be used as a mouse to control the cursor in the PDA. This allows us to centralize in the PDA the control of domotic devices using the UPnP domotic service. For instance, browsing a few buttons using the joystick and then sending the selected button is a less demanding solution for UPnP communications. Note that this implementation is not incompatible with the possibility of using the monitoring service in the system. The domotic devices could activate an alarm if the wheelchair's user presses an emergency button in the wheelchair's console, for example. In figure 8, the architecture of this modification is shown.

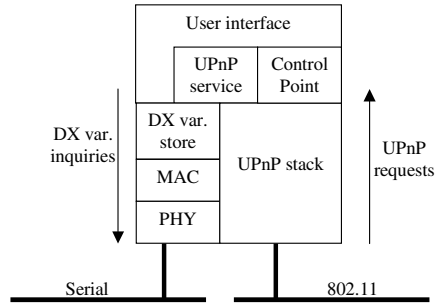


Fig. 8. Modified structure of the software in the PDA

Another aspect that has to be taken into account is that the use of Bluetooth to communicate the DX Bridge to the PDA may be problematic due to possible interferences. First, due to the coexistence between Bluetooth and 802.11b links in the same device, Bluetooth 1.2 links are recommended because they implement *adaptive frequency hopping* (AFH) techniques that allow coexistence with 802.11 links [16]. AFH works by using fewer than 79 channels in the frequency hopping mechanism if the Bluetooth device detects that there is interference on some of these frequencies. In this way, frequencies occupied by 802.11 are avoided, allowing co-existence.

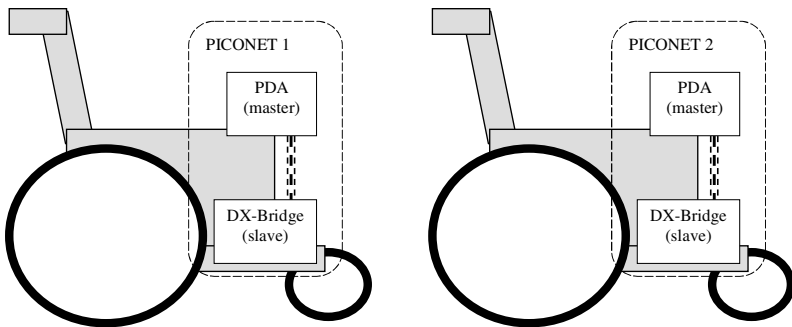
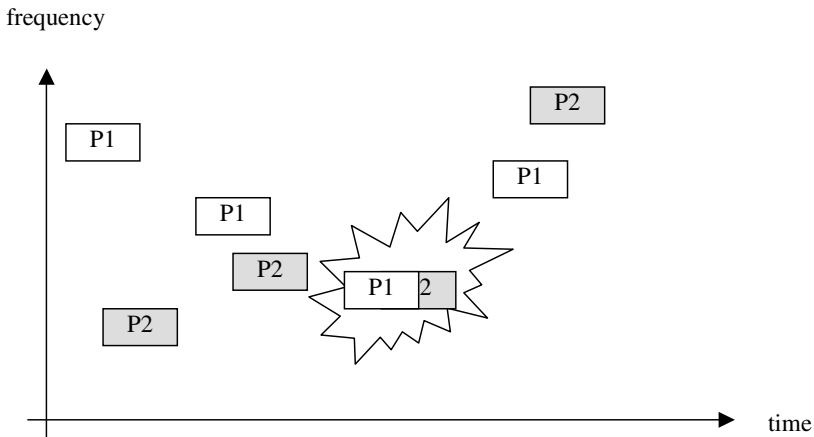


Fig. 9. Independent Bluetooth links formed in wheelchairs

In addition, the Bluetooth link may also suffer from other independent Bluetooth devices in the same area. The wheelchair may be placed in an area where other users have active Bluetooth links, for instance mobile phones, PDAs, MP3 players with headphones, or even other wheelchairs equipped with the same devices. In this case, every independent Bluetooth device uses its own frequency hopping sequence, so collisions may occur if two or more of these devices happen to choose the same frequency. If the DX variables sent through the Bluetooth link have time constraints (for instance, if they are used to activate an alarm or to control an external device) then some QoS guarantees are needed. For instance, in [24] the worst-case deadline failure probability of Bluetooth messages is obtained as a function of the number of independent interfering devices. Depending on the application, a number as low as 5-10 (which may not be strange in conference halls, airports, etc.) may be unacceptable.




**Fig. 10.** Example of collision between piconets

## 6 Conclusions

In this paper, we describe how an electric wheelchair is adapted to allow its integration in AmI Systems. The wheelchair used is a prototype from a previous project, which is based on a DX Bus, a *de facto* standard in electric wheelchairs. We describe a new DX hardware module that allows the DX sub-systems to communicate with other devices, particularly with the user interface (in our prototype, a StrongARM®-based PDA equipped with serial and 802.11b interfaces). We also describe the software developed for this PDA, which is based on the UPnP architecture. The current prototype is able to interact with external devices in two ways: on the one hand, commands sent by the wheelchair UCM (joystick) are interpreted as commands to control a domotic system or the like. On the other hand, DX variables can be accessed from external devices, both as an UPnP device, as well as via a simple web page.

**Version 1.0**

Wheelchair's gateway:  Monitoring service.

---

**Wheelchair's status:**

| Variable         | Value | Variable   | Value |
|------------------|-------|------------|-------|
| DX communication | OK    |            |       |
| UCM Status       | 0     | PM Status  | 0     |
| Mode             | 0     | Profile    | 0     |
| Joystick X       | 0     | Joystick Y | 0     |
| Battery          | 0     | Switches   | 0     |

---

**Actions:**

Variable ID:  Value:

Previous action results: No action requested.

---

**Fig. 11.** Web page of the UPnP monitoring service

## Acknowledgments

The research presented in this paper has been developed within the project *Heterorred* "Study and development of a heterogeneous personal area network for interoperability and access to wireless services and communications", funded by the Spanish Ministry of Science and Technology under grant No. TIC2001-1868-C03.

## References

- [1] ISTAG; Scenarios for Ambient Intelligence in 2010; Final Report, Feb 2001, EC 2001: <http://www.cordis.lu/ist/istag.htm>
- [2] J.L. Sevillano et al.: On the Design of Ambient Intelligent Systems in the Context of Assistive Technologies. 9th International Conference on Computers Helping People with Special Needs, Paris 2004. LNCS 3118, pp. 914-921. Springer 2004.
- [3] A. Civit, J. Abascal: "TetraNauta: A Wheelchair Controller for Users with Very Severe Mobility Restrictions". Improving the Quality of Life for the European Citizen. I. Plasencia, E. Ballabio (eds.). pp. 336-341. IOS Press, 1998.
- [4] S. Vicente, et al.: "TetraNauta: a intelligent wheelchair for users with very severe mobility restrictions". Proc. IEEE Int. Conf. on Control Applications, Pp: 778 - 783. Sept. 2002.
- [5] S. Vicente Díaz. "Una aportación al guiado de sillas de ruedas eléctricas en entornos estructurados" (in spanish). PhD Thesis. Universidad de Sevilla, July 2001.
- [6] <http://www.dynamic-controls.co.nz>
- [7] Mike Meade, "DX Key Technical Description. For DX Key Application Designers." Dynamic Controls Ltd., 1997.
- [8] <http://www.silabs.com>
- [9] <http://www.ehsa.com/>
- [10] J. Abascal, J.L. Sevillano, A. Civit, G. Jiménez, J. Falcó: Integration of heterogeneous networks to support the application of Ambient Intelligence in assistive environments. IFIP Conf. on Home Oriented Informatics & Telematics HOIT 2005 (York, U.K.. April 2005).
- [11] Jeronimo M, West J: UPnP Design by Example: A software developer's guide to Universal Plug and Play. Intel Press (2003).
- [12] <http://www.intel.com/labs/connectivity/upnp/index.htm>
- [13] <http://www.plug-n-play-technologies.com/>
- [14] 802.11 Working Group's Site: <http://grouper.ieee.org/groups/802.11/>
- [15] R. Casas, "Sistema interoperable de localización en interiores aplicado a tecnología asistencial" (in spanish). PhD Thesis. Universidad de Zaragoza, Spain. Sept. 2004.
- [16] The official Bluetooth website. <http://www.bluetooth.com>
- [17] Axel Lankenau, Thomas Röfer: "A versatile and safe mobility assistant", IEEE Robotics and automation magazine, p29-37, March 2001.
- [18] Dan Ding, R. A. Cooper: "Electric-powered wheelchairs: a review of current technology and insight into future directions". IEEE Control Systems Magazine, p22-34, April 2005.
- [19] <http://www.atc.us.es/?op=investigacion>
- [20] J.C. Haartsen, The Bluetooth Radio System., IEEE Personal Communications 7(2000) 28-36.
- [21] The Bluetooth Special Interest Group: Specification Of Bluetooth System - Core Vol.1 V1.1. Feb 2001. [www.bluetooth.com](http://www.bluetooth.com).
- [22] D. Cascado, J.L. Sevillano, S. Vicente, F. Díaz del Río, G. Jiménez, A. Linares, A. Civit-Balcells. Modeling Effects of Co-channel Interference over Performance in Single-Slave Bluetooth Piconets. The 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2004).
- [23] D. Cascado. Study and evaluation of a wireless communication system for personal area networks (in spanish). Ph. D. Tesis. University of Seville (2003).
- [24] J.L. Sevillano, D. Cascado, F. Díaz del Río, S. Vicente, G. Jiménez, A. Civit-Balcells. Statistical QoS guarantees in Bluetooth under co-channel interference. 10th IFIP International Conference on Personal Wireless Communications (PWC 2005).

- [25] R.C. Simpson and S.P. Levine. Automatic adaptation in the NavChair assistive wheelchair navigation system. *IEEE Trans. Rehab. Eng.*, vol. 7, no. 4, pp. 452–463, 1999.
- [26] Nisbet, P.D. (2002) Assessment and Training of Children for Powered Mobility in the UK. *Technology & Disability* 14 (2002). p173–182. IOS Press. ISSN 1055-4181/02.
- [27] Thierry Ernst. E-Wheelchair: A Communication System Based on IPv6 and NEMO. 2nd International Conference On Smart homes and health Telematic (ICOST2004).
- [28] Chuan-Heng Hsiao et al. A design of small-area automatic wheelchair. *IEEE International Conference on Networking, Sensing & Control*, p1341-1345. (Taiwan 2004)
- [29] R.A. Cooper, M.L. Boninger, R. Cooper, and A.R. Dobson. Technical perspectives: Use of the Independence 3000 iBOT Transporter at home and in the community. *J. Spinal Cord Med.*, vol. 26, no. 1, pp. 79–85, 2003.
- [30] Salvador, Z., Bonail, B., Lafuente, A., Larrea, M., Abascal, J. and Gardeazabal, L., AmIChair: Ambient Intelligence and Intelligent Wheelchairs. *Proceedings of HOIT 2005 (Home Oriented Informatics and Telematics 2005)*.
- [31] H. Hoyer. The OMNI wheelchair. *Service Robot: An International Journal*, Vol.1 No.1, MCB University Press Limited, Bradford, England, pp. 26-29, 1995.

# Collaborative Discovery Through Biological Language Modeling Interface

Madhavi Ganapathiraju<sup>1</sup>, Vijayalaxmi Manoharan<sup>1,2</sup>, Raj Reddy<sup>1</sup>,  
and Judith Klein-Seetharaman<sup>1,2</sup>

<sup>1</sup>Language Technologies Institute, Carnegie Mellon University, Pittsburgh USA

<sup>2</sup>Department of Structural Biology, University of Pittsburgh, Pittsburgh USA  
{madhavi, viji, rr, judithks}@cs.cmu.edu

**Abstract.** Scientific progress is exponentially increasing, and a typical example is the progress in the area of computational biology. Here, problems pertaining to biology and biochemistry are being solved by way of analogy through the application of computational theories from physics, mathematics, statistical mechanics, material science and computer science. More recently, theories from language processing have been applied to the mapping of protein sequences to their structure, dynamics and function under the Biological Language Modeling project. Scientists from diverse computational and linguistics backgrounds collaborate with experimental biologists and have made significant scientific contributions. The essential component of this collaborative discovery is the web server of the biological language modeling toolkit that enables the computational and non-computational scientists to interface and collaborate with each other. The web server acts as the computational laboratory to which researchers from a variety of scientific disciplines and geographical locations come to characterize specific attributes pertaining to their protein or groups of proteins of interest using the available tools. They then combine the results with their domain expertise to arrive at conclusions. The web server is also useful for education of students entering into the research field in computational biology in general. In this paper, we describe this web server and the results that were arrived at through local and global collaboration and education.

## 1 Introduction

Technological evolution is an exponential process [1]. A key characteristic of scientific and technological advancement is interdisciplinary research and collaborative discovery. These are believed to be major sources for innovation because unsolved problems in one area can potentially benefit from technologies borrowed from other areas of research [2]. On the other hand, disjoint study of the unsolved problems by experts from disjoint fields can lead to cognitive artifacts [3].

The shift from individual research to collaborative discovery is accelerated by free and open access to information, tools and services on the Internet. The Open Source Initiative (OSI) has been created in 1998 to promote the open source concept whereby a software product is made freely available under a public license, so that programmers



can read, modify and redistribute the software—a process by which the software evolves [4]. Thus, software often becomes more reliable, scalable, secure and better performing, primarily due to world wide review and contribution. The concept is commercially successful, often open source software holding a better market share than proprietary software [5]. Through web-dissemination, the ease of accessing publicly funded research also constantly increases (“Open Research”) and Open Access to scholarly journal articles is extending. Thus, Open Source, Open Research and Open Access together become major forces driving and supporting research globally towards collaborative discovery. Collaborative efforts are generally viewed as beneficial because the researchers in a collaboration benefit from each others’ perspectives and expertise, and researchers may ask uncommon questions and draw untypical hypotheses in their collaborators’ domain as a result of their different domain specializations [6].

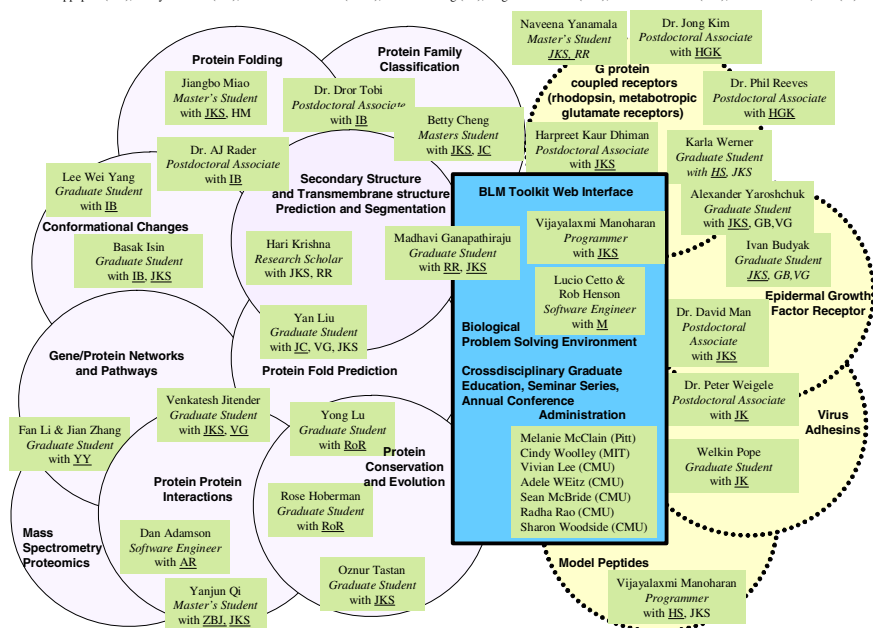
In this paper, we will present the Biological Language Modeling (BLM) project as an example of a project involved in collaborative discovery, which brings together researchers from diverse backgrounds, ranging from biological chemistry to various specializations in computational sciences, with particular focus on language technologies [7, 8]. The goal of the BLM project is to identify and address specific unsolved problems in biological sequence analysis that are amenable to the application of expertise gained in the area of natural language processing [9]. Experimental data generated by biologists is used by the computational researchers to draw inferences, which are in turn validated and verified by the experimental biologists. Figure 1 gives an overview of the BLM project (at the time of its creation in 2002); the computational focus areas are shown on the left and experimental focus areas are shown on the right (see legend). Since the collaborators belong to geographically dispersed institutions, methods of interaction between researchers include email, video conferences, individual and conference calls, local discussion groups, journal clubs and seminars. National and international workshops and conferences are also employed in this project. In addition, a web service has been created and continues to evolve to facilitate learning, collaboration, inference and discovery in the BLM project [10].

In the following, we describe the biological language modeling project web interface, the tools provided and how they aid collaborative research and education. We present the specific analogy between language and biology as a case study to illustrate convergence of technologies. Such convergence is not limited to the analogy between language and biology, and can encompass other technologies and/or types of data. Numerous results achieved through local and global collaboration between scientists in experimental and computational fields are presented. The web interface has also been used for education in research, and the utility is demonstrated through the scientific results achieved through such use.

## 2 Biological Language Modeling

The complexity of biological data requires sophisticated algorithms and approaches for its organization, processing, analysis, visualization and retrieval. Biological data, like many other types of data, is stored in large databases and diverse websites. For

Faculty: Raj Reddy (RR), Judith Klein-Seetharaman (JKS), Ivet Bahar (IB), Jaime Carbonell (JC), Yiming Yang (YY), Roni Rosenfeld (RoR), Vanathi Gopalakrishnan (VG), Alain Rappaport (AR), Cathy Costello (CC), H. Gobind Khorana (HGK), Jonathan King (JK), Hagai Meirovitch (HM), Michele Loewen (ML), The Mathworks, Inc. (M)



**Fig. 1. Concept map of the projects and people associated with the BLM project.** Solid circles (left) indicate the biological research areas investigated using computational language technologies. Dotted circles (right) indicate the experimental laboratory systems that are used to test computational models or to generate data for the development of computational models. The blue square box (center) shows the education and outreach activities. Faculty names are abbreviated as indicated in the box on top and are referred to in the boxes representing students, postdocs, programmers and software engineers. The primary advisors are underlined. Only PI's of the projects are listed as advisors with full names, others only by their abbreviations. For further details visit [www.cs.cmu.edu/~blmt](http://www.cs.cmu.edu/~blmt).

example, web servers focused on gathering and annotating specific types of biological data include the PDB [11], Swissprot [12], Ensembl [13], PFAM [14], GPCRDB[15], to name just a few. Innumerable websites publish specialized datasets. Similarly, there is a vast range and diversity of tools developed to analyze the gathered data for the purpose of a specific prediction task such as secondary structure prediction or protein family classification. Since generally a given biological problem requires the use of several data sources and tools, a number of servers exist that combine the most popular ones. Most prominent amongst them are the Biology workbench [16-19], EXPASY [20, 21] and NCBI [22, 23]. Furthermore, new methods for a variety of prediction tasks in the biological domain are being developed at a fast pace.

Analysis of human language data faces similar challenges in terms of size, complexity and diversity. Therefore, one general approach to the treatment of biological data has been the application of language technologies [9]. Language and biological data share the mapping process of encoded data such as text or speech to meaning, analogous to mapping sequences to the functions of their encoded proteins,

for example. Language is analyzed with two conceptually different approaches (1) linguistic, rule-driven, that aims at revealing the fundamental mechanism of the mapping process or (2) data-driven, where practical approaches to a given task such as information retrieval in a search engine are sought, irrespective of the underlying mechanism. Both approaches have been applied to the analysis of biological data [9, 24-27]. Note, that the use of language-based approaches meant here is different from the use of language technologies in information extraction from biological literature [28-31] and not from biological data itself. Similar to the language area, the data-driven approach has a broader range of applicability due to the availability of large amounts of data. Thus, statistical and machine learning approaches are essential to computational biology, but the specific application to language provides a rich source of intuitive understanding and detail that is directly transferable to the biology domain due to the inherent similarity between biology and language. These methods have therefore been the foundation for several recent publications in the computational biology domain. Thus, protein domain boundary prediction [32], protein family prediction [33, 34], transmembrane helix recognition [35], protein fold prediction [36], protein secondary structure prediction [36, 37], feature extraction and comparison of biological sequences [38-40], identification of evolutionary conservation pressures [41] and protein-protein interaction prediction [42] have all used methods borrowed from language technologies.

The utilization of methods borrowed from language technologies applied to biological discovery in the BLM project has led to the development of several tools based on data-driven approaches to language. The tools from the biology/language case study described here are available on the Biological Language Modeling Toolkit (BLMT) web server. The BLMT web server provides means for the analysis of biological data complementary to the existing suites of available tools by providing improvements over existing methods, tackling new prediction tasks, including new data sources and extracting features from biological data differently from previous approaches, as described below. Users need to upload or choose an input sequence, a protein Swiss-Prot ID or set of sequences or IDs of interest. Specific tools or a set of tools are applied as entered by the user who can then download/view the computed results. The visualization of the results is a major component of the web-interface and is designed to encourage creative discovery and biological insight. In addition to the tools, specialized datasets that have been compiled from various publicly available databases are available at <http://flan.blm.cs.cmu.edu/>.

## 3 Tools

### 3.1 N-gram Extraction and Analysis

Searching for a substring from large text data is a problem in various areas of computer science that has been dealt with efficiently using efficient data structures like suffix trees [43] and suffix arrays [44]. A sub-string in language is referred to as N-gram, a linear string of words with N being the number of words in the string. Usually words are spelled in one defined way, and the words are used in one specific sequence, therefore the concept of N-gram computations defines searches in a text

database for exact matches. This defines an important difference to biological sequence data, in which “approximate” matches such as amino acids can be replaced by similar amino acids without changing the structure or function of the protein. Therefore, traditional sequence analysis approaches have made use of sequence alignments in which a matrix is used to identify matching sequences while simultaneously allowing for substitutions in the aligned sequences. The wide-spread use of N-grams in language applications, however, suggests that N-grams may also have utility as features in biological applications. This was recently demonstrated by use of N-grams for the classification of proteins, where N-gram features were shown to carry complementary information to sequence alignments [33, 34]. Therefore, it is likely that N-grams may also be useful features for other prediction tasks in computational biology.

The input for the N-gram extraction tool on the BLMT web server (Fig. 2) is a sequence file of any length (there is essentially no upper length limit due to the efficiency of the data storage method) and any encoding. Protein sequence and whole genome DNA files are supported. Re-encoding of such sequences by any of the alphabets stored on our server, or provided as an input by the user, is also supported. In each case, the output is a computation of n-gram statistics in the sequence provided, or with respect to a reference sequence for comparison. String matching using the N-gram concept is fastest when using data structures such as suffix trees and arrays, since these are optimized for large data and fast processing requirements and have been used in the bioinformatics domain for a variety of applications, for example whole genome sequence alignment [45]. The BLMT web server takes a whole-genome file as input and generates suffix arrays, Least Common Prefix (LCP) array [46] and Rank arrays for the file. The generated binary data could be stored for later use or statistical data of the genome can be computed. Also, pre-generated arrays of selected Genome datasets from the NCBI are available for download. The output is provided as text or in visual form, for example, where the location of n-grams that are rare (or significant) and found only in a given protein and not in any other related dataset (for example, the whole genome).

The N-gram statistics calculated by the tool are meant to be used as input for other applications. For example, the “N-gram comparison” interface on the BLMT server provides options for uploading whole Genome files and comparing their genome signatures. N-grams are generated from a specified reference file and its frequency of occurrence across the compared genome files is computed. Options for evaluating the statistical significance of the findings are the inclusion of randomized datasets and the calculation of means and standard deviations of n-gram frequencies for given sequence datasets based on the difference between expected and observed n-gram frequencies. An in depth description of the applications built on the n-gram statistical tool and its uses can be found in reference [47]. The strongest evidence for the utility of N-gram statistics comes from its use in protein family classification [33, 34]. Here, N-gram feature vectors derived from different sub-classes are used as input into classification algorithms, such as Naïve Bayes or Decision Trees. In conjunction with chi-square feature selection, a technique wide-spread in the language domain, it was shown that these features allow more accurate classification than Support Vector Machines or other methods including BLAST in conjunction with alignment-based features [48].

**BLMT - N-gram Extraction - Microsoft Internet Explorer**

**A** Upload Genome files [Format](#)

31 >g1|10582006|gb|AAG20661.1| Vng2626h **B** jbaa  
 32 >g1|10581935|gb|AAG20600.1| Vng2544h [na]obaa  
 33 >g1|10581474|gb|AAG20207.1| Vng2049c [Haloba  
 33 >g1|10579667|gb|AAG18659.1| Vng0019h [Haloba  
 34 >g1|10581346|gb|AAG20097.1| Vng1904h [Haloba  
 34 >g1|10579684|gb|AAG18674.1| Vng0034h [Haloba  
 35 >g1|10581809|gb|AAG20493.1| Vng2402h [Haloba  
 35 >g1|10580410|gb|AAG19293.1| Vng0840h [Haloba  
 36 >g1|10581543|gb|AAG20268.1| Vng2129h [Haloba

Enter the n-gram length *\*required*  to

**Display Options:**

Display counts of only top  n-grams

Display counts of all ngrams

Display n-grams with count

Sort By *\**  By Alphabet  By Count

*\* Sorting is only by count if the top few ngram counts option is chosen.*

**Other Statistical Data:**

Total Number of Proteins

Average length of Proteins

Length of Proteins

Check here if you want to display protein header with the length.

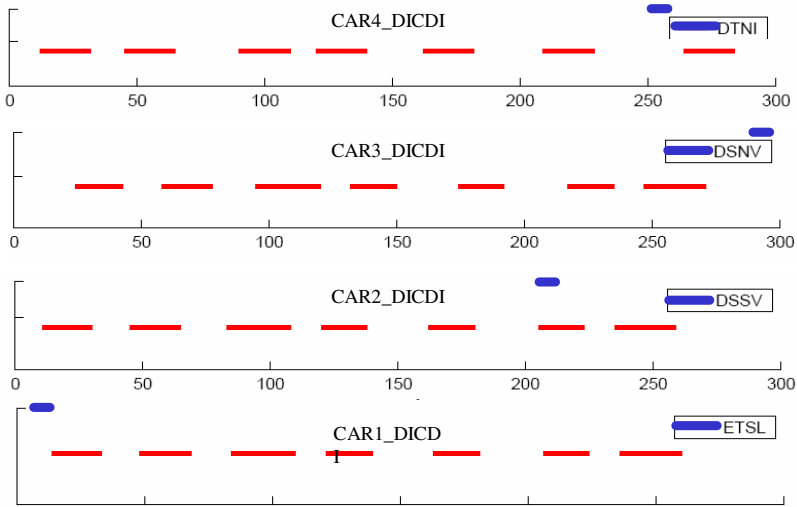
**C**

|        |    |
|--------|----|
| AAAAAA | :  |
| LSGGE  | 19 |
| GSGKT  | 18 |
| GTGKT  | 17 |
| EALEA  | 16 |
| GADAV  | 16 |
| CPVDA  | 15 |
| RVKNN  | 15 |
| GKTTL  | 15 |
| ELLER  | 15 |
| EEALE  | 15 |
| ALEAG  | 14 |
| VKNNL  | 14 |
| PGT GK | 14 |
| LRELG  | 14 |

**Fig. 2.** N-gram Extraction and Analysis on the BLMT web server. Parameters marked with red \* are mandatory. One can either upload Genome Files from the local computer or choose a file from the dataset provided on the server (not shown in the figure). A basic usage would be an input file and the n-gram which could be entered as a range. The output would be a list of the frequency of occurrences of the n-grams in the genome. To display the n-grams along with their frequency (as shown in inset C) one has to check the option “display n-grams with count”. B is an example output file giving length of proteins in a genome file with header information. This file is generated on choosing ‘Length of Proteins’ option under “Other Statistical Data”. C is an example output for 5-grams and their frequency of occurrence in the genome file sorted by frequency.

### 3.2 N-gram and Regular Expression Visualization

The utility of N-gram features for accurate protein family classification demonstrated that there is biological meaning in N-grams. Since the N-gram features that are most predictive for a given class can be analyzed, the analogy between language and biology in this context allows interpreting the computational results from a biological



**Fig. 3.** Location of the PDZ regular expression [DE][ST]x[LVMI] in class E G Protein coupled receptors. x indicates presence of any amino acid and square brackets presence of either of the amino acids inside them. The red lines mark the helix regions in the sequences and the blue lines mark the occurrence of the regular expression in that sequence. The legend gives the actual motif in the protein sequence.

perspective. This is done by inspecting the location of highly predictive N-gram features with respect to its position in the protein sequence of interest. Thus, we were able to show that the important N-grams correlate with motifs of known functional importance, such as enzyme active sites and protein interface motifs [33]. In some cases, N-gram motifs were overlapping and had substitutions in them. To incorporate such cases, these overlapping N-gram motifs can also be combined into regular expression patterns. To visualize N-grams and regular expression patterns, the BLMT web server provides visual capabilities to identify motif occurrences in a chosen sequence or set of sequences or Swiss-Prot IDs. For example, Fig. 3 shows a known function protein interaction motif, the PDZ motif, [DE][ST]x[LVMI], with respect to its occurrence in Class E G protein coupled receptors. The location of the motif is indicated with simultaneous annotation of secondary structure elements. For the G protein coupled receptor family, these are the locations of the transmembrane helices, shown as red bars in Fig. 3. The secondary structure information used for this annotation can be extracted if the data is provided in Swiss Prot format or the Swiss-Prot ID of the protein has been specified. The N-gram motifs can also be visualized with respect to their position in a three-dimensional structure, as described for association measures, below (Fig. 5).

**3.3 Association Measures: Yule’s Q-Statistic, Mutual Information**

Association and correlation measures and distance metrics are important for any data-driven discipline. Two association measures, typically used in language applications, Yules’ Q statistic and Mutual Information, are beginning to be explored for

applications in biological sequence analysis. For example, mutual information has been used in nucleic acid structure prediction [49, 50], differentiating coding and non-coding regions in DNA [51] and in gene clustering [52] and to identify functional building blocks in protein sequences [38].

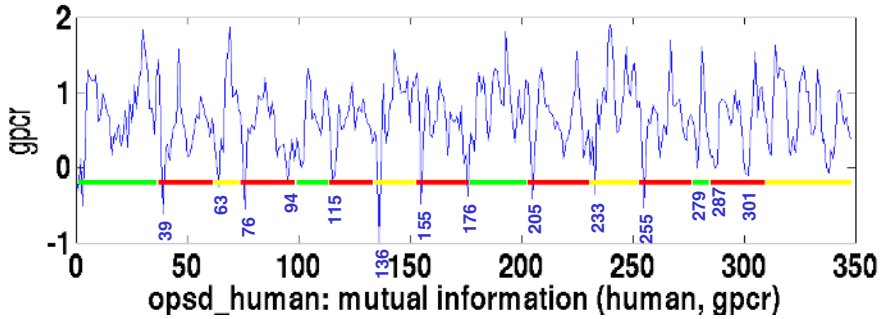
Yule's Q statistic indicates presence or absence of correlation between two variables by value between -1 and 1. A positive value implies that the variables are positively correlated. Likewise, negatively correlated variables have a negative Yule value. The tool computes a 20x20 amino acid Yule table from the uploaded input protein sequence and the training set. Subsequently, the user can analyze this Yule table, or plot the Yule values along a protein sequence. The training set is recommended to contain at least 60,000 amino acids [40].

Mutual Information has been used successfully for protein sequence segmentation in single protein sequences, given a set of protein sequences as reference input. Within a window of size four, the association between the occurrences of amino acids is calculated and is given as a function of amino acid sequence position. Minima are identified and used to predict feature boundaries based on a collection of protein sequences [38]. The biological significance lies in applications such as chimera construction between two members of a given family, where precise break-points in the sequences suitable for joining complementary fragments from two related sequences need to be identified. An example is shown in Fig. 4. Mutual Information was calculated using the G protein coupled receptor family and plotted along the rhodopsin sequence. A comparison with the locations of transmembrane helices shows that there is a strong correlation between the breakpoints and the helix ends. Importantly, no information such as hydrophobicity or assumptions/expert knowledge was entered to obtain this plot. The data is also shown as a color mapping onto the rhodopsin structure in Fig. 5. The BLMT web server uses chime software [53] for this purpose. The advantage of this means of visualization is the ability to identify correlations between sequence breakpoints in the three-dimensional structure.

### 3.4 Positional Property Conservation

As described above, for the complementary nature of N-grams to traditional sequence alignment-based features, the traditional bioinformatics view of sequence conservation can be limited. The analogy to language allows further means for complementing this traditional view. Sequence conservation is useful when a position is highly conserved. However, at intermediate conservation levels, biological insight from multiple sequence alignments could be achieved if the conserved underlying property, rather than the identity of a given amino acid, could be identified. Two language-based methods have been developed to examine the conservation of amino acid properties with respect to their positions in a sequence in order to complement sequence conservation methods [41]. One method is implemented using a Gaussian distribution to model property conservation, while the second uses variance to identify conservation patterns. Given a multiple sequence alignment (MSA), a protein sequence and a property file, the properties of amino acids at the alignment positions is examined in relation with the corresponding position of the amino acid in the protein sequence of interest. Application of the positional property conservation to the rhodopsin sequence is shown in Fig. 6. The conservation of helix-packing using

the scale described in ref. [54] in the alignment of class A G protein coupled receptors is plotted along the rhodopsin sequence. Strong correlation between the locations of the helices and the locations of conservation of helix packing moment can be seen.



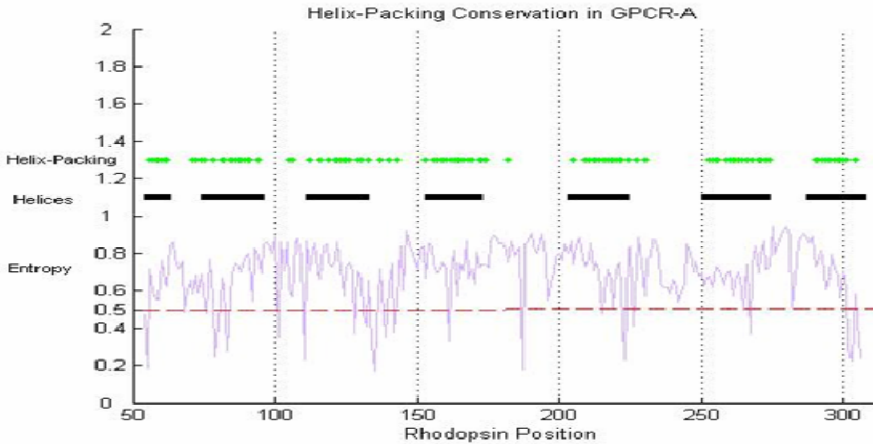
**Fig. 4.** A. Mutual information derived from the family of G protein coupled receptor sequences as the dataset with respect to the rhodopsin sequence (swiss prot id opsd\_human). The horizontal line indicates the extracellular (green), helices (red) and cytoplasmic (yellow) regions of the protein. The blue labels indicate the positions of the predicted breakpoints.



**Fig. 5.** Mapping of the Mutual Information values shown in Fig. 4 onto the rhodopsin crystal structure (PDB Id 1F88). High Mutual Information values are denoted by blue and low values by red color.

Again, visualization is critical in the analysis of such results. Mapping of amino acid property correlation to a sequence or structure can help derive new biological hypotheses. Given a set of properties, the BLMT web server can generate text, bar graphs, line graphs and matrix graphs of the correlations between a set of related properties.





**Fig. 6.** Conservation of a property, here helix-packing, in a multiple sequence alignment, here GPCR-A, mapped along a single member of the family, here the rhodopsin sequence. Sites with conservation levels suitable for property conservation analysis are defined as having normalized standard deviation below 0.8 and entropy above 0.5, are indicated as dots.

### 3.5 Protein-Protein Interaction Prediction Tool

Finally, the analogy between biological data and language data can be taken a step further. In the previous sections, the analogy was literal, because biological sequence data is “written” using the amino acid and nucleotide alphabets and its derivatives. However, we recently also used the analogy to consider the equivalent of a word, not a sequence motif or N-gram, but instead an entire protein without referring to its sequence. At that level, a number of equivalents for “meaning” can be derived, for example the function of a protein or its position within a protein interaction network, to name just a few. We focused on the equivalent of “meaning” as “protein interaction partner”. Now, the features used to predict if a pair of proteins is interacting can be, but are not restricted to, sequence-based features. A new supervised classifier borrowing implementation details from language technologies has been developed to combine multiple biological evidence to predict protein-protein interaction in yeast, including microarray expression data, protein function data, cellular localization data and others. A random forest algorithm [55] was used to measure similarity between proteins followed by K nearest neighbor algorithm for classification of interacting protein pairs [42]. The tool was tested on the yeast pheromone response pathway to classify 300 potential interacting pairs [42, 56]. 70.45% of the predictions were true, while only 4.55% were false positives. The high performance is attributed to the algorithm and the choice of features. Included as features were not only the datasets that specifically look for protein interaction, but also indirect information such as mRNA expression data, protein-DNA binding data, and protein localization data. The BLMT web server allows retrieving prediction scores of 2 or more proteins and probable interacting partners for a given protein. The user can either enter a yeast protein ORF or choose it from the list provided. A threshold prediction score has to be entered for the second option.

### 3.6 Compute All

The BLMT web server is designed for researchers with varying expertise and backgrounds. Thus, individual tools are supported with maximal flexibility for modifications in the parameters. For researchers new to the biological language modeling field and/or project, we also implemented a “compute all” mode. This mode is designed to allow for a general overview of the tools provided and a generic application of all tools to a specific user-defined biological problem. In this mode, a given protein and set of reference proteins can be entered, and all of the tools are applied. The user is provided options to further tune parameters of the specified tool and generate results as the user’s understanding of the results increases.

### 3.7 Visualizations

The goal of using the analogy is to provide a platform for biological knowledge discovery. Therefore, visualization of statistical data is a major component of the web-interface as it greatly aids interpretation of results. The web service provides three ways of visualizing the data. (1) Numerical output is provided as a text file and can be used as a feature for further analysis. (2) The computed data are plotted as 2D graphs using Matlab [57]. Examples are shown in Figs. 3, 4 and 5. The Swiss-Prot format of the protein sequence allows plotting the secondary structure positions of the protein for reference. (3) When PDB files are available, computed values are plotted on the 3D structure of the protein. All three types of visualization complement each other in helping the formulation of biological hypotheses. Visualization of results is what acts the central binding point of the collaborators of different backgrounds. The web server separates the computation and visualization into different modules, thereby allowing researchers to feed in newer algorithms into the system and allowing biologists to use them for analysis of their sequences in comparison to different experimental information (for example, Swissprot data giving segment level information or PDB data giving 3D structural information).

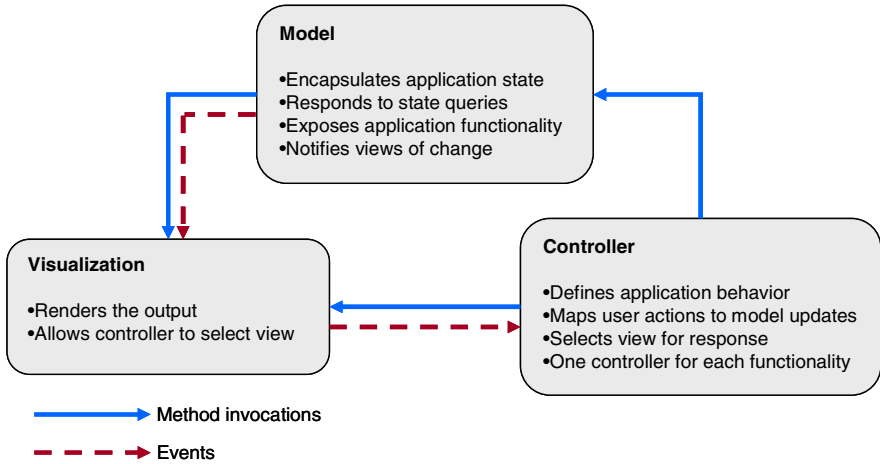
## 4 Software Architecture

The software architecture of the BLMT follows the standard architectural pattern of the model-view-controller (MVC) (see Fig. 7) [58].

The tools (models) are independent from visualization and user choices, but they communicate with each other through sharing of data and events. Biological sequence data are often stored in different formats, so the tools have been designed in a flexible way to allow input in any of the standard sequence formats. Processing of these data yields numerical or textual results that form the input for the visualization. The controller acts as the interface between models and the user (see a snapshot of the interface for the n-gram tool in Fig. 8). There is a different controller for each tool.

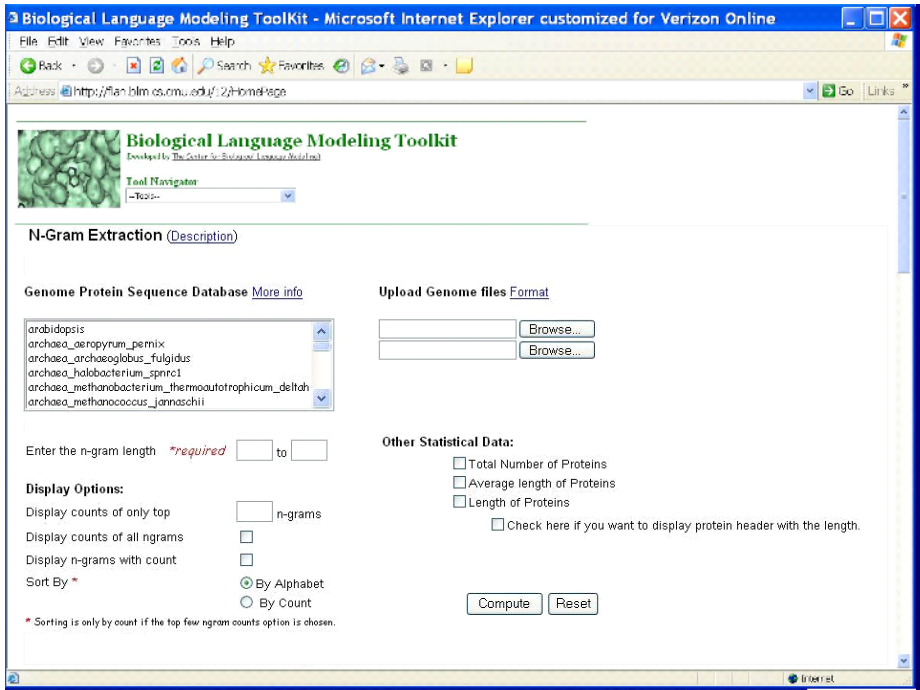
The controller, or web interface input page for the tool, accepts the user parameters, input sequences and output formats provided or chosen by the user. The models are the tools applied to the data as initiated by the controller to yield results for the sequence input, and the output is generated in textual or numeric format. The

visualization module is called by the controller. The visualization module reads the output, reformats it into the appropriate format needed for visualization. If the mode of the visualization needs to be changed, the controller will send the appropriate event. The visualization module also allows the user to change the view as per user request and alerts the controller, which invokes the model methods again with new parameters.



**Fig. 7.** Model View Controller software architecture that describes the BLMT web interface. The tools and visualization modules are independent of each other, but they may use common data and/or output from other modules. For each tool, there is a web interface component that acts as its controller. It takes the user inputs and invokes the execution of the tools. The output parameters are set by the controller based on input or choices selected by the users. The visualization module is invoked by the controller, and reads the data generated by the model, reformats it appropriately for each type of visualization and returns the display to the user. The three components interact with each other only through the data and events that are passed between them, and this architecture therefore allows the different tools to easily integrate with the ability for updates in any of the components without disruption of the system. This shields the user from operational details. New tools for analysis and visualization can be seamlessly integrated into the system, allowing it to evolve continuously.

The separation of the three modules, namely models, view and control, allow the BLMT and its web interface to evolve with the development of new tools and new visualizations. For example, the addition of new tools such as position specific property conservation and Yule's Q-statistic computation has been seamlessly integrated into the previously existing interface. The MVC architecture is best suited for this collaborative discovery setting, wherein different tools and visualization methods are continuously added to the system by different computational researchers. The burden of understanding the usage of the new tools and visualizations is kept minimal—the controller presents users with a consistent interface that they are familiar with, thereby allowing them to use the new features easily.



**Fig. 8.** Web interface snapshot: A snapshot of the web interface for the N-gram extraction tool is shown (this forms the “controller” as discussed in the software architecture section). The interface allows the users to choose preloaded sequences files or to upload sequences by users. The value of ‘n’ to compute the n-grams is taken as the input parameter. Output options can also be chosen here by the users (number of most frequent n-grams, whether to sort the output by counts or alphabetically). After the input and output parameters are given, the compute button is pressed, which invokes the model component of the MVC architecture and in turn, the visualization component of the MVC. The controller then waits for modifications of the Visualization options by the user (another controller), or for the invocation of the controller of another tool.

## 5 BLMT Web Interface and Collaborative Discovery

### 5.1 Local Collaboration

Traditional computational work is done at statistical level—the data sets (from any chosen application area) are analyzed and inferences are drawn on how the models perform statistically. For example, in case of speech recognition, it is of interest to see how a new algorithm can reduce the word error rate on a collection of sentences spoken by speakers—it is rarely of interest to evaluate a recognition engine on a very specific sentence. In the case of computational biology this holds true, since algorithms are also designed and optimized for statistical performance. However, a major difference is the application of given tools to specific proteins or genes. An entire biology lab usually studies one such protein by experimental methods, and

when available, would use computational tools to draw inferences about that specific protein. While this scenario would be unusual in the language arena, it is the typical case in biological chemistry.

With this in mind, the BLMT web interface provides many of the computational tools developed as part of the BLMT project for use by its biologist partners and outside users. A large number of results have been achieved by this collaboration, and are briefly listed below with references to representative published articles. (For a full list of references, see the publications link on the BLMT-website [59].)

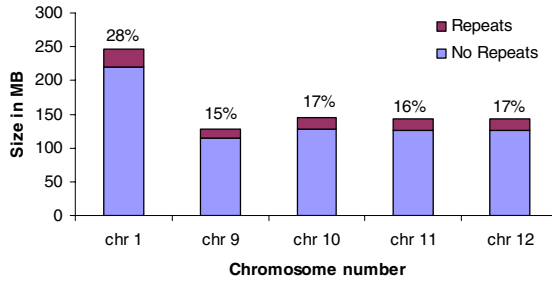
- N-gram analysis of whole genomes and determination of genome signatures [39, 60].
- Secondary structure and transmembrane structure prediction using latent semantic analysis analogy and context sensitive vocabulary [61-63]
- Transmembrane helix characterization using natural language word association measures [40]
- Text-inspired algorithms for transmembrane helix segmentation [35]
- Protein classification based on text classification techniques [48, 64]

## 5.2 Global Collaboration and Education

The web server is also being used for analysis of protein sequences by researchers outside of the biological language modeling research group [65] and for education. Courses taught to graduate students in bioinformatics at the International Institute of Information Technology, India, use the tools on the web server to analyze, interpret and understand the correlations between statistical characteristics of biological sequences with respect to structural characteristics of proteins. For example, in the course “*Research projects in computational biology*” taught by one of the authors (MG), the students performed protein sequence analysis with the n-gram tools, to study their use in gene recognition, prediction of circular permutations in proteins, tandem repeats and phylogenetic analysis. The tools are used to reaffirm previously known results so that the students experience themselves how biological hypotheses can be created and tested using bioinformatics approaches. Below, we give three examples of the utility of the web server for such educational purposes.

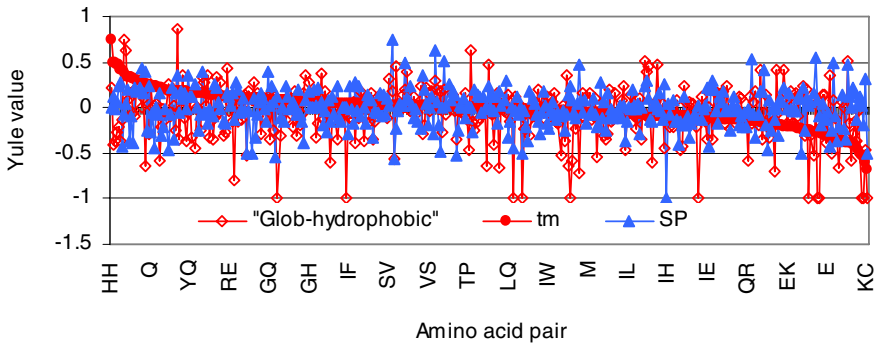
**Student project 1: Scalable algorithm for recognition of variable number tandem repeats (VNTRs)** (new algorithm built over BLMT by Sunaina Reddy, and Vijayalakshmi Sundararajan, graduate students at International Institute of Information Technology, Hyderabad, India): Genome sequences contain regions of repetitive sequences. The repeated sequences vary in length, ranging from short (2-7 nucleotides long) to very long (1000 nucleotides long) sequences. Identification of repeat regions is important because the number of tandem repeats varies among individuals, and predisposes them for diseases. However, identification of repeat regions is challenging because they contain many nucleotide substitutions. Algorithms exist to detect short tandem repeats in sequences, but they are limited by the length of the repeated sequence or by the number of substitutions they can handle.

Student project 1 was to develop a new algorithm to detect tandem repeats using the functionalities of the BLMT [47]. The suffix array data structure of the BLMT groups identical sequences that are located dispersed throughout the genome. With



**Fig. 9.** Percentage of chromosomal regions consisting of tandem repeats. Human chromosomes have been studied with the tandem repeat finding algorithm developed over BLMT. 5 of the human chromosomes studied are lined along the x-axis, and their size is shown along the y-axis. Each chromosome is shown in split color to indicate non-repetitive regions (blue) and repetitive regions (maroon), and percentage of repetitive region is shown in the labels above.

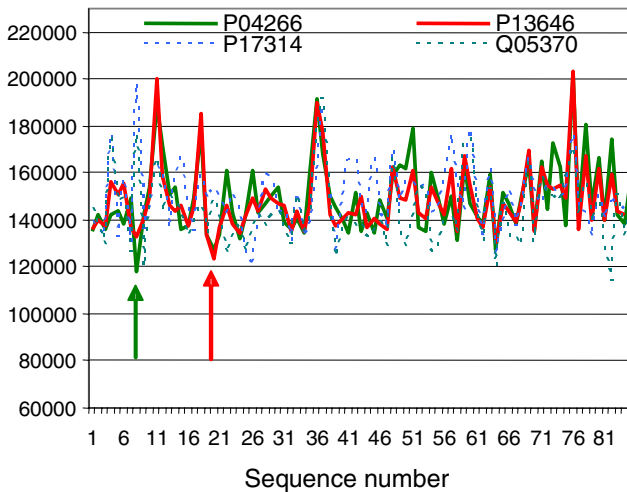
the help of the longest common prefix array and the rank array, the lengths and positions of repeats are computed. The tandem repeat finder algorithm scans through the suffix and LCP arrays to find the lengths of repeated sequences, and uses the rank array to find whether the positions of these repeated sequences are in tandem. Since the algorithm is built over efficient data structures used in the BLMT, it is scalable to recognition over genomic data sizes, and has been used to estimate percentage of repeat sequences in chromosomes (see Fig. 9). This tool has been built over the open source BLMT software.



**Fig. 10.** Yule’s Q-statistic applied to hydrophobic helices from transmembrane (red, solid circles) and globular (red, open circles) proteins and to signal peptides (blue triangles). The Yule values are computed for these three data sets using the BLMT web server. The outputs are generated in plain text format, and are reprocessed to generate a comparative graph. The amino acid pairs are arranged along the x-axis such that they occur in descending order of their Yule value in TM helices. It can be seen that for a given pair of amino acids, its Yule value (y-axis) is distinct for each data set. This suggests that the preferred neighborhood of amino acids is different in each of these data sets.

**Student project 2: Characterization of hydrophobic helices from globular and transmembrane proteins** (new results established with the use of BLMT by Pamela Misra (intern at Indian Institute of Science): Transmembrane (TM) helix prediction algorithms often incorrectly predict globular helices and signal peptide sequences to be of TM type. The goal of student project 2 was to identify if correlations between amino acids in globular helices, SP sequences and actual TM regions differ. Yule's Q-statistic association measure has been used previously to quantify preferences of amino acid pairs to occur along different faces of a helix [40]. Hydrophobic segments in globular proteins that are potentially confusable as TM segments were obtained using the Kyte-Doolittle method [66]. Signal peptide and TM data have been taken from [67]. Yule's Q-statistic was computed using the BLMT for the three data sets (see Fig. 10). The results show that Yule values vary between the three data sets and may prove useful features for TM prediction algorithms.

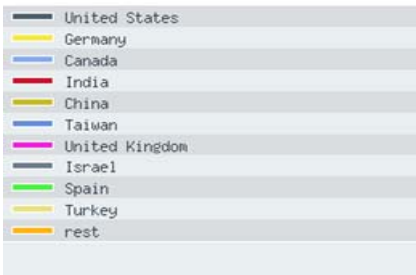
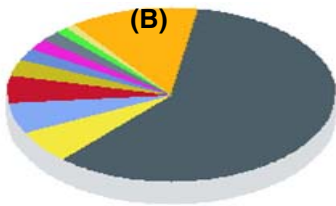
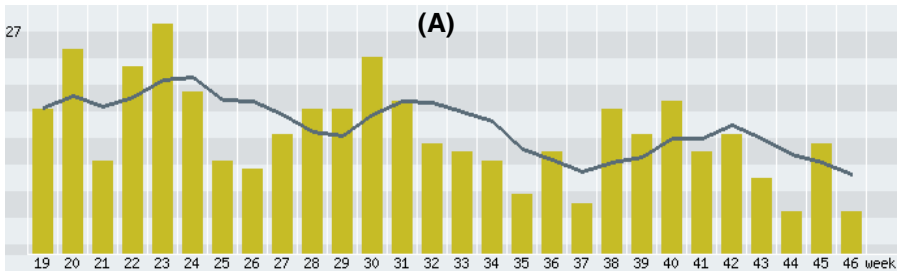
**Student project 3: Detection of circularly permuted proteins using language modeling** (new algorithm built over BLMT by Viswatej Nalubolu, graduate student at International Institute of Information Technology, Hyderabad, India): Circular



**Fig. 11.** Plots to identify circularly permuted pairs of proteins: Each point on the x-axis corresponds to a specific protein in the database. Perplexity of generating a protein with the language models of all the proteins being considered are computed and plotted against those proteins. For example, a point (6, 140000) for protein N indicates that the perplexity of protein N with respect to the language model of protein 6 is 140000. The lower the perplexity, the greater is the similarity between the two sequences. Perplexity plot for 4 sequences (accession numbers P04266, P13646, P17314 and Q05370 in Swissprot) are shown. The first two sequences are circular permutations of each other. Points 8 and 20 on x-axis, where the arrows are located, correspond to these two proteins, and the fact that they are circular permutations of each other is marked by a dip in perplexity with respect to themselves and also with respect to their circular permutation partners in the plot. It can also be seen the plots of these two sequences are very similar and have a correlation coefficient  $> 0.85$ .

permutations in proteins are defined as a genetic segment swapping event in which the N and C-termini of a protein are fused and a break is made elsewhere to generate new N and C termini of the circularly permuted protein [68]. In student project 3, an algorithm has been built over the BLMT to identify pairs of proteins that are circular permutations of each other. As a first step, proteins that have high sequence similarity (>30% identity) are filtered out using BLAST. N-gram models of all the proteins are built using the BLMT.

The perplexity of generating each protein with respect to every other protein is computed with respect to this model for n-grams of length 4, with back-off to lengths 3, 2 and 1. Potential circular permutation pairs are identified such that the perplexity of one protein with respect to another is very close to the perplexity of the protein with itself. This narrows down the potential circular permutation partners to a few sequences for each protein. When perplexity plots are computed as shown in Fig. 11, the correlation coefficient between plots of circular permutation pairs is > 0.85. A second step analysis with dot-matrix comparison or other known circular permutation finding algorithms may then be used to confirm the exact pairs.



**Fig. 12.** (A) Number of accesses to the web server per week. There have so far been about 2000 hits to the web page since its launch in 2002. (B) Access to the web server by country. The web server is employed by global users.



## 6 Utility of the Web Interface

The tools offered via the BLMT web interface are locally available for users of the biological language modeling group. Users with a computational background may use either the BLMT source code or the web interface, but the experimental biologists who are typically from different institutions and geographical locations use exclusively the web interface for their work. The web interface allows users to choose tools and parameters, to use preloaded data sequences or upload their own sequences, and to visualize the results in graphs or color coded renderings, or to download numerical data to use as input to other sequences. The usage statistics of the web server indicate regular access to the web server (Fig. 12A) by globally distributed researchers (Fig. 12B).

## 7 Availability

The Biological Language Modeling project web page is located at: <http://www.cs.cmu.edu/~blmt>. The BLMT web interface which is also linked from the project webpage is at: <http://flan.blm.cs.cmu.edu/>. For those who are interested in the open source, the underlying tools are available on the web server for download.

## 8 Summary

The successful application of the analogy between language and biology to several prediction tasks in computational biology demonstrates that the convergence of technologies is an important method to advance science by extrapolating from one discipline to another. The use of analogies is particularly useful because it bridges the communication gap between seemingly unrelated disciplines. This allows efficient exploitation of parallel advances in each separate field. Therefore, researchers with varying expertise and backgrounds can now work cooperatively on a common problem. To facilitate such collaborative research, individual tools need to be supported with maximal flexibility for modifications in the parameters. Here, we have described an interactive web server, the BLMT web server, which provides tools derived from language technologies for the analysis, storage, processing, retrieval and interpretation of biological data. For a researcher entering a new field, e.g. a particular protein, all of the methods developed can be applied simultaneously, allowing a high chance for generating new biological hypotheses. On the BLMT web server, the user is provided options to tune parameters. Finally, the BLMT web interface forms the ground where the computational and experimental researchers exchange their research: tools are developed based on interactions with biologists, and are made available on the web. The tools are then validated by experimental data or are used to infer new hypotheses. Different computational tools overlap and enhance each other – by sharing features or by one tool providing as output the input for another tool. The web interface alleviates the problem of familiarity with background implementation and presents the user with the view of the results that are of concern, rather than with the details on how to handle the code. On the other hand, the open source availability

of the code allows the evolution of the algorithms from the benefit of the World Wide Web review in addition to the classical peer-review system.

## Acknowledgments

The work described here was in part supported by NSF grants 0225656, 0225636 and CAREER CC044917, NIH - NLM grant 1R01LM007994 and the Sofya Kovalevskaya Prize from the Humboldt - Foundation / Zukunftsinvestitionsprogramm der Bundesregierung Deutschland.

## References

- [1] Kurzweil, R., *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. 2000: Penguin. 400.
- [2] Klein-Seetharaman, J. and R. Reddy, Biological Language Modeling: Convergence of computational linguistics and biological chemistry, in *Converging Technologies for Improving Human Performance*. Nanotechnology, Biotechnology, Information Technology and Cognitive Science, W.S. Bainbridge, Editor. 2002, National Science Foundation: Arlington, Virginia. p. 378-385.
- [3] Jones, P.H. and C.P. Nemeth, Cognitive artefacts in complex work. *Lecture Notes in Artificial Intelligence*, 2005. LNAI/LNCS 3345: p. 152-83.
- [4] OSI, Open Source Initiative: <http://www.opensource.org/>.
- [5] Wheeler, D.A., *Why Open Source Software / Free Software (OSS/FS, FLOSS, or FOSS)? Look at the Numbers!* 2005.
- [6] Okada, T. and H.A. Simon, Collaborative discovery in a scientific domain. *Cognitive Science: A Multidisciplinary Journal*, 1997. 21(2): p. 109-146.
- [7] Klein-Seetharaman, J. and R. Reddy. Biological Language Modeling: Convergence of Computational Linguistics and Biological Chemistry. in NSF Workshop "Converging Technology (NBIC) for Improving Human Performance. 2002.
- [8] Klein-Seetharaman, J. The Use of Analogies for Interdisciplinary Research in the Convergence of Nano-, Bio- and Information Technology. in *NSF Report on Societal Implications of Nanoscience and Nanotechnology*. 2005.
- [9] Ganapathiraju, M., et al., Computational Biology and Language. *Lecture notes in artificial intelligence*, 2005. LNCS/LNAI 3345: p. 25-47.
- [10] Manoharan, V., M. Ganapathiraju, and J. Klein-Seetharaman. BLMT Web Server: Interactive Language Technologies for Analogous Biological Data. in *Workshop on Ambient Intelligence and (Everyday) Life*. 2005. San-Sebastian, Spain.
- [11] Berman, H.M., et al., The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*, 2000. 7 Suppl: p. 957-9.
- [12] Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res*, 1999. 27(1): p. 49-54.
- [13] Hubbard, T., et al., Ensembl 2005. *Nucleic Acids Res*, 2005. 33(Database issue): p. D447-53.
- [14] Bateman, A., et al., The Pfam protein families database. *Nucleic Acids Res*, 2002. 30(1): p. 276-280.
- [15] Horn, D.L., et al., Why have group A streptococci remained susceptible to penicillin? Report on a symposium. *Clin Infect Dis*, 1998. 26(6): p. 1341-5.

- [16] Subramaniam, S., The Biology Workbench--a seamless database and analysis environment for the biologist. *Proteins*, 1998. 32(1): p. 1-2.
- [17] Sauro, H.M., et al., Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *Omics*, 2003. 7(4): p. 355-72.
- [18] Biology-WorkBench. <http://bsw-uiuc.net/>.
- [19] Systems-Biology-WorkBench. <http://workbench.sdsc.edu/>.
- [20] Gasteiger, E., et al., ExpASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 2003. 31(13): p. 3784-8.
- [21] ExpASy. <http://www.expasy.org/>.
- [22] Jenuth, J.P., The NCBI. Publicly available tools and resources on the Web. *Methods Mol Biol*, 2000. 132: p. 301-12.
- [23] NCBI. <http://www.ncbi.nlm.nih.gov/>.
- [24] Searls, D.B. and M.O. Noordewier, Pattern-matching search of DNA sequences using logic grammars, in *Proceedings of the 7th Conference on Artificial Intelligence Applications*. 1991, IEEE. p. 3-9.
- [25] Searls, D.B., The language of genes. *Nature*, 2002. 420(6912): p. 211-7.
- [26] Bolshoy, A., et al., Enhancement of the nucleosomal pattern in sequences of lower complexity. *Nucl. Acids. Res.*, 1997. 25(16): p. 3248-3254.
- [27] Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 1997. 268(1): p. 78-94.
- [28] Hearst, M., Untangling Text Data Mining, in *37th Annual Meeting of the Association for Computer Linguistics*. 1999: College Park, MD, USA. p. 3-10.
- [29] Pustejovsky, J., et al., Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations, in *Pacific Symposium on Biocomputing*. 2002: Hawaii, USA. p. 362-73.
- [30] Friedman, C., et al., GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, in *Bioinformatics*. 2001. p. S74-82.
- [31] Hatzivassiloglou, V., P.A. Duboue, and A. Rzhetsky, Disambiguating proteins, genes, and RNA in text: a machine learning approach, in *Bioinformatics*. 2001. p. S97-106.
- [32] Coin, L., A. Bateman, and R. Durbin, Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl. Acad. Sci. USA*, 2003. 100(8): p. 4516-4520.
- [33] Vries, J., et al., A Sequence Alignment-Independent Method For Protein Classification, in *Applied Bioinformatics*. 2004. p. 137-48.
- [34] Cheng, B., J. Carbonell, and J. Klein-Seetharaman, Protein Classification based on Text Document Classification Techniques, in *Proteins - Structure, Function and Bioinformatics*. 2005. p. 955-70.
- [35] Cheng, B., J. Carbonell, and J. Klein-Seetharaman, A Machine Text-Inspired Machine Learning Approach for Identification of Transmembrane Helix Boundaries, in *15th International Symposium on Methodologies for Intelligent Systems*, Saratoga Springs. 2004: New York, USA. p. 29-37.
- [36] Liu, Y., et al., Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics*, 2004. 20(17): p. 3099-3107.
- [37] Ganapathiraju, M., et al., Characterization of protein secondary structure using latent semantic analysis. *IEEE Signal Processing magazine*, 2004. 21(3): p. 78-87.
- [38] Weissner, D. and J. Klein-Seetharaman, Identification of Fundamental Building Blocks in Protein Sequences Using Statistical Association Measures, in *ACM Symposium on Applied Computing*. 2004: Nicosia, Cyprus. p. 154-161.

- [39] Ganapathiraju, M., et al. Comparative n-gram analysis of whole-genome sequences. in HLT2002: Human Language Technologies Conference. 2002. San Diego, USA.
- [40] Ganapathiraju, M., et al. Yule value tables from protein datasets of different categories: emphasis on transmembrane proteins. in SCI2004: Eighth World Multi-Conference on Systemics, Cybernetics and Informatics. 2004. Orlando, Florida, USA.
- [41] Hoberman, R., J. Klein-Seetharaman, and R. Rosenfeld, Inferring Property Selection Pressure from Positional Residue Conservation. *Applied Bioinformatics*, 2004. 3(2-3): p. 167-180.
- [42] Qi, Y., J. Klein-Seetharaman, and Z. Bar-Joseph, Random forest similarity for protein-protein interaction prediction from multiple sources, in 10th Pacific Symposium on Biocomputing. 2005: Hawaii. p. 531-542.
- [43] Weiner, P., Linear pattern matching algorithms, in In: Proc. of the 14th Annual Symp. on Switching and Automata Theory. 1973: University of Iowa. p. 1-11.
- [44] Manber, U. and G. Meyers, A new method for on-line string searches. *SIAM Journal on Computing*, 1993. 22(5): p. 935-948.
- [45] Delcher, A.L., et al., Alignment of whole genomes, in *Nucleic Acids Res.* 1999. p. 2369-76.
- [46] Kasai, T., et al. Linear-Time Longest-Common-Prefix computation in Suffix Arrays and Its applications. in Annual Symposium on Combinatorial Pattern Matching CPM-2001. 2001. Jerusalem, Israel,.
- [47] Ganapathiraju, M., V. Manoharan, and J. Klein-Seetharaman, BLMT: Statistical Sequence Analysis using N-grams, in *J. Applied Bioinformatics*. 2004. p. 193-200.
- [48] Cheng, B., J. Carbonell, and J. Klein-Seetharaman, Protein Classification based on Text Document Classification Techniques. *Proteins - Structure, Function and Bioinformatics*, 2005. 58(4): p. 955-70.
- [49] Chiu, D.K. and T. Kolodziejczak, Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci*, 1991. 7(3): p. 347-52.
- [50] Akmaev, V.R., S.T. Kelley, and G.D. Stormo, Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, 2000. 16(6): p. 501-12.
- [51] Grosse, I., et al., Average mutual information of coding and noncoding DNA. *Pac Symp Biocomput*, 2000: p. 614-23.
- [52] Butte, A.J. and I.S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 2000: p. 418-29.
- [53] <http://www.mdlchime.com/>.
- [54] Liu, W., et al., Helix packing moments reveal diversity and conservation in membrane protein structure. *J Mol Biol*, 2004. 337(3): p. 713-29.
- [55] Breiman, L. Random forests. in *Machine Learning*. 2001.
- [56] Qi, Y., Z. Bar-Joseph, and J. Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction, in *Proteins - Structure, Function and Bioinformatics*. 2005.
- [57] MATLAB. <http://www.mathworks.com/>.
- [58] Shaw, M. and D. Garlan, *Software Architecture: Perspectives on an Emerging Discipline*. 1006: Prentice Hall.
- [59] BLMT-Publications. <http://www.cs.cmu.edu/~blmt/publications.html>. 2005.
- [60] Klein-Seetharaman, J., et al. Rare and frequent amino acid n-grams in whole-genome protein sequences. in RECOMB'02: The Sixth Annual International Conference on Research in Computational Molecular Biology. 2002. Washington, USA.
- [61] Ganapathiraju, M., et al., Characterization of protein secondary structure using latent semantic analysis, in *IEEE Signal Processing magazine*. 2004. p. 78-87.

- [62] Ganapathiraju, M., et al., Computational Biology and Language, in Lecture notes in artificial intelligence. 2005. p. 25-47.
- [63] Liu, Y., et al., Context Sensitive Vocabulary And its Application in Protein Secondary Structure Prediction, in ACM SIGIR Conference. 2004. p. 538-9.
- [64] Qi, Y., Z. Bar-Joseph, and J. Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins - Structure, Function and Bioinformatics*, 2005. in press.
- [65] Dong, Q.W., X.L. Wang, and L. Lin. N-gram Statistics and Linguistic Features Analysis of Whole Genome Protein Sequences. in HUPPO 3rd Annual World Congress. 2004. Beijing, China.
- [66] Kyte, J. and R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, in *J Mol Biol*. 1982. p. 105-32.
- [67] Chen, C.P., A. Kernytsky, and B. Rost, Transmembrane helix predictions revisited, in *Protein Sci*. 2002. p. 2774-91.
- [68] Uliel, S., et al., A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, 1999. 15(11): p. 930-6.

# Author Index

- Almeida, Aitor 239  
Amaya, Claudio 285  
Anthony, Caroline 178  
Aube, Julien 227
- Baskett, Michael 256  
Bonanni, Leonardo 130
- Cai, Yang 67  
Cascado, Daniel 285  
Civit-Balcells, Antón 285  
Cukierman, Diana 178
- de Vries, Arjen P. 103
- England, David 256  
Esnaola, Urko 198  
Fernández, Javier 239  
Ganapathiraju, Madhavi 300  
Garcia, Daniel 239  
García, Iván 239  
Jiménez, Gabriel 285  
Kimura, Atsunobu 142  
Klein-Seetharaman, Judith 300  
Kobayashi, Minoru 142  
Kwan, Chiman 227
- Lenat, Douglas 1  
Linares, Alejandro 285  
Llewellyn-Jones, David 256  
López-de-Ipiña, Diego 239
- Manoharan, Vijayalaxmi 300  
Matuszek, Cynthia 1  
Maurer, Uwe 86
- Mei, Gang 227  
Minakuchi, Mitsuru 157
- Nakamura, Satoshi 157
- Pantic, Maja 32  
Panton, Kathy 1  
Peters, Geoffrey 178
- Reddy, Raj 300  
Ren, ZuBing 227  
Rochet, Cedrick 227  
Rowe, Anthony 86
- Sáinz, David 239  
Sainz Salces, Fausto J. 256  
Schneider, Dave 1  
Schwartz, Michael 178  
Sevillano, J. Luis 285  
Shepard, Blake 1  
Shimada, Yoshihiro 142  
Siegel, Nick 1  
Siewiorek, Daniel 86  
Smailagic, Asim 86  
Smithers, Tim 198  
Stanford, Vincent 227
- Tanaka, Katsumi 157
- van Doorn, Mark 103  
Vázquez, Juan Ignacio 239  
Vicente, Saturnino 285
- Witbrock, Michael 1
- Xu, Roger 227