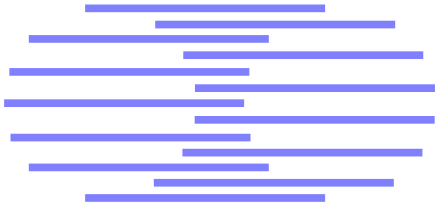
IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse



# Activity Report 1999

IDIAP–Com 2000-01

February 2000

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP

Martigny - Valais - Suisse

# Institut Dalle Molle d'Intelligence Artificielle Perceptive

## MEMBERS

Supporting:

- Swiss Confederation, Federal Office for Education and Science (FOES)
- State of Valais
- City of Martigny
- Swisscom

Affiliated:

- Swiss Federal Institute of Technology at Lausanne (EPFL)
- University of Geneva

## FOUNDATION COUNCIL

Pierre Crittin (Chairman, President of the City of Martigny), Jean-Pierre Rausis (Secretary, Director of BERSY), Hervé Bourlard (Director of IDIAP, Professor at EPFL), Pierre Dal Pont (Director of NOFIDA), Daniel Forchelet (Swisscom, Skill Family Manager), Gilbert Fournier (State of Valais), Jurg Hérold (Director of CIMO SA), Nicolas Markwalder (Attorney at Law, Delegate of the Economic Commission, Bern), Jérôme Sierro (University of Geneva), Dominique de Werra (Professor, Vice-President of EPFL).

## BOARD OF DIRECTORS

Jean-Pierre Rausis (Chairman, Director of BERSY), Pierre Dal Pont (Secretary, Director of NOFIDA), Hervé Bourlard (Director of IDIAP, Professor at EPFL), Daniel Forchelet (Swisscom, Skill Family Manager), Gilbert Fournier (State of Valais), Jurg Hérold (Director CIMO SA), Nicolas Markwalder (Attorney at Law, Delegate of the Economic Commission, Bern), Christian Pellegrini (Professor, University of Geneva), Léopold Pflug (Professor, EPFL).

## SCIENTIFIC COMMITTEE

Prof. Christian Pellegrini (Chairman, University of Geneva, CH), Prof. Hervé Bourlard (Director IDIAP, Professor EPFL), Dr. Robin Breckenridge (F. Hofmann-La Roche Ltd, CH), Prof. Giovanni Coray (EPFL, CH), Dr. J. Cywinsky (Institute of Medical Technology, CH), Prof. Wulfram Gerstner (EPFL, CH), Prof. Martin Hasler (EPFL, CH), Prof. Jean-Paul Haton (CRIN/INRIA, F), Prof. Beat Hirsbrunner (University of Fribourg, CH), Prof. Rolf Ingold (University of Fribourg, CH), Prof. Eric Keller (University of Lausanne, CH), Prof. Nelson Morgan (ICSI and UCB, Berkeley, USA), Prof. Beat Pfister (ETH, CH), Prof. Thierry Pun (University of Geneva, CH), Prof. Ian Smith (EPFL, CH), Mr. Robert Van Kommer (Swisscom, CH), Prof. Eric Vittoz (CSEM and EPFL, CH), Prof. Christian Wellekens (EURECOM, F).

# Table of Contents

# 1 General Overview of the Institute

## 1.1 Introduction

The Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP, "Institut Dalle Molle d'Intelligence Artificielle Perceptive", `http://www.idiap.ch`) is a semi-private non-profit research institute founded in 1991 to celebrate the 20th anniversary of the Dalle Molle Foundation. It is the third research centre initiated by the Dalle Molle Foundation, after ISSCO in Geneva (`http://www.issco.ch`) and IDSIA in Lugano (`http://www.idsia.ch`).

In November 1996, and as initially planned at the establishment of the institute, IDIAP acquired the status of Research Foundation (IDIAP Foundation), now independent of the Dalle Molle Foundation, counting as founders the City of Martigny, the State of Valais, the Swiss Federal Institute of Technology in Lausanne (EPFL), the University of Geneva and Swisscom.

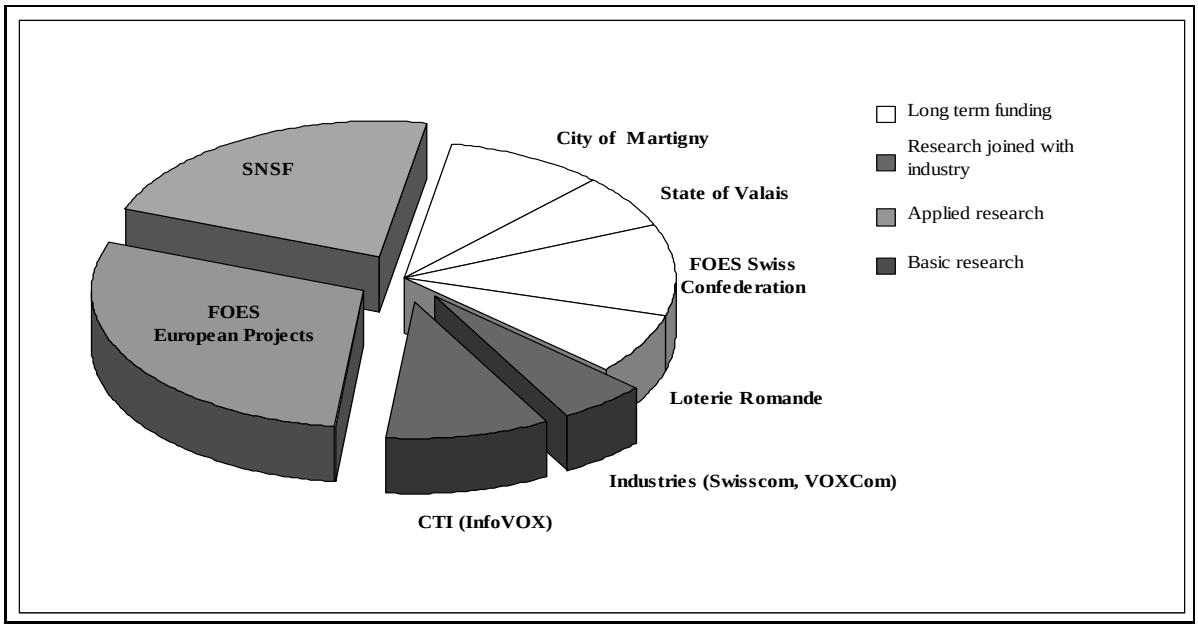



Figure 1: *Relative distribution of IDIAP funding in 1999.*

Today, IDIAP is primarily funded by long-term support from the the Swiss Confederation (Federal Office for Education and Science, FOES), the State of Valais, the City of Martigny, and Swisscom. The "Loterie Romande" is also supporting our research efforts with annual grants. In addition, IDIAP receives substantial research grants from the Swiss National Science Foundation (SNSF) for basic research projects (mainly for PhD students), and from FOES in the framework of European projects. The relative distribution of IDIAP's funding in 1999 is illustrated in Figure 1.

For the last few years, there has been an average of about 25-30 scientists in residence at IDIAP including permanent staff, postdoctoral fellows, PhD students, and short-medium term visitors.

The general management structure of IDIAP is illustrated in Figure 2 and is composed of a Foundation Council, a Board of Directors, and a Scientific Committee (advising the Board of Directors). In the framework of VOXCom, a spin-off company of IDIAP, a small Economic Relations Committee

has also been set up, which is responsible for publicizing IDIAP's research results across the industrial world, as well as providing IDIAP with new research opportunities of particular interest to industry.

The activities carried out at IDIAP can be described as follows: research and development activities, participation in European and national research projects, collaborations with organizations and companies, and teaching and training activities. IDIAP's mission therefore consists in:

- Carrying out fundamental and applied research activities aiming at long and medium term industrial transfer.
- Teaching and training activities.

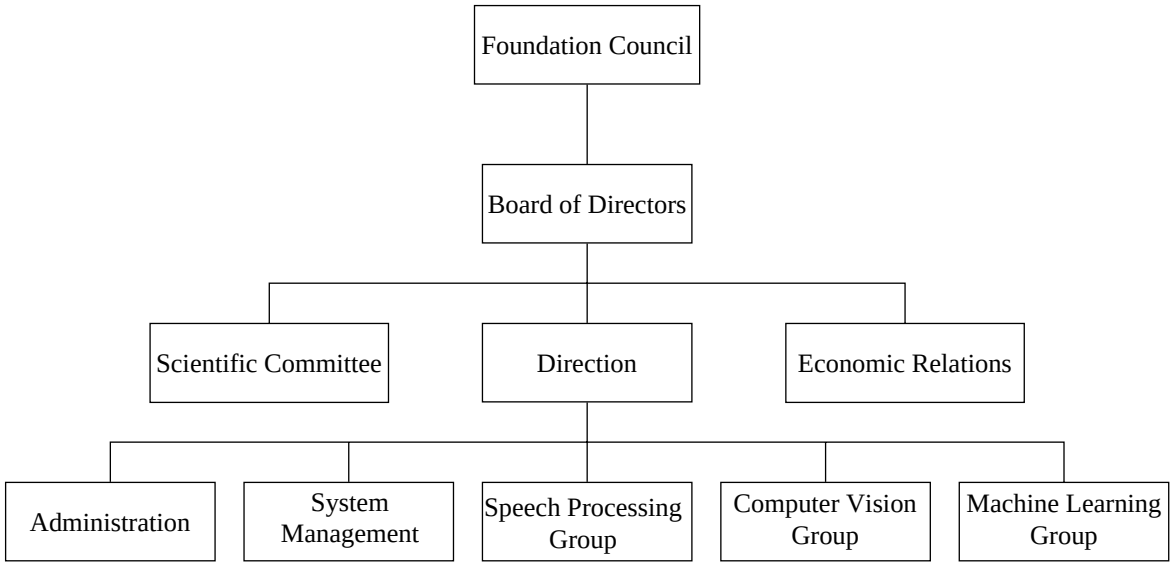



Figure 2: *IDIAP Structure.*

In 1999, IDIAP's activities have continued to flourish. The number of national and international projects has significantly increased, and many new projects (as briefly discussed below) were granted to IDIAP in 1999. Also, the partnerships with academic institutions have significantly grown (see more about this below). Moreover, thanks to the continued support of our authorities, and to our most competent personnel, motivated to the highest level, IDIAP is recognized as a highly sought partner in the well defined areas (speech processing, computer vision, machine learning) they decided to focus on. It is now our job to continue to concentrate our research and development activities on those areas, while fostering technology transfer through industrial partnerships.

As further discussed below, another important success for IDIAP in 1999 was its pre-selection as one of the potential "National Centres of Competence in Research", which could start on Jan. 1, 2001. Following a very strict and competitive selection process (based on national and international reviews), IDIAP was indeed shortlisted (with two other proposals from ETHZ) as the proposal with the highest chances of funding. This project, entitled "Interactive Multimodal Information Management (IM)2", and centered on many of the research activities of IDIAP, was submitted in collaboration with many national (EPFL, ETHZ, Univ. of Geneva, Univ. of Fribourg, Univ. of Bern) and international (ICSI in Berkeley, and Eurecom in Sophia-Antipolis) organizations. While tightening further the links between IDIAP and many of its university partners, this project would also confirm the leading role of IDIAP at the national level in the targeted research areas. Final selection will take place in 2000, based on a full proposal to be submitted by March 15, 2000.

## 1.2 Research and Development Activities

The main 1999 research activities of IDIAP, which focus on medium and long term objectives, are described in detail in the present Activity Report. Focusing on a few, well defined, research axes, along the general theme of **multimodal interaction**, IDIAP carries out fundamental research and develops prototype systems (to validate its research results) along three complementary directions:

- **Speech Processing, including all aspects of automatic speech recognition and speaker verification**.

  This involves the development and testing of advanced and state-of-the-art speech recognition systems (ranging from small to large vocabularies, from speaker dependent to speaker independent, from isolated words to continuous speech and keyword/keyphrase spotting). While mainly focusing on telephone speech, this work is also applied to microphone input. Current research activities mainly focus on improving speech unit models towards better robustness to noise and speaking styles. Amongst other activities, this involves online adaptation techniques, further developments to hidden Markov models (HMM) and hybrid systems using HMMs together with artificial neural networks, as well as advanced research in sub-band and multi-stream processing (as pioneered by IDIAP, together with the Faculté Polytechnique de Mons in Belgium, the International Computer Science Institute in Berkeley, USA, and the Oregon Graduate Institute in Portland, USA). Large vocabulary speech recognition systems, involving complex pronunciation dictionaries and rules, as well as advanced grammatical constraints, are also developed and tested.

  As described below, the IDIAP Speech Processing group is involved in numerous national and international projects (such as the European ESPRIT, ACTS, COST, and TMR projects).

  In speaker verification, most of the research activities so far have focused on the improvement of current state-of-the-art algorithms, and on the development of innovative solutions combining concurrent and/or complementary strategies. These last two years, IDIAP participated in the international NIST (National Institute of Standards and Technology, USA) evaluation and showed that their technology was at the leading edge in that field.

  The main applications and prototype systems which have been developed and tested so far were oriented towards: advanced interactive voice servers (e.g., for accessing remote databases), personal call assistants, calling card applications (involving voice dialing and speaker verification), automatic audio indexing and retrieval, and multimodal (speech and vision) user verification systems.

  Finally, to facilitate research, as well as multi-lingual system development, IDIAP is also actively involved in speech database collection, labeling, and management activities. These activities have mainly taken place in the framework of our cooperation with Swisscom (Polyphone and GSM data), or as part of a large European effort towards the development of large multi-lingual databases.

- **Computer Vision, including object recognition, motion analysis, sensor fusion, and document recognition**.

  Computer Vision in general deals with the automatic analysis and interpretation of visual scenes. Although this is a very broad field, the strategy of the group is to focus on research topics that are driven by specific target applications, with the aim of developing new technologies in the area of multimodal interfaces, access security, and information management and retrieval.

  Through activities in various projects, the group has acquired expertise in the areas of object detection and recognition, shape analysis and representation, motion analysis and recognition, sensor fusion, and document analysis and recognition.

Much of this work benefits from collaboration with the speech recognition group (e.g., in motion recognition using HMMs or in multimodal recognition using multi-stream processing) and the machine learning group (e.g., classifier combination, support vector machines).

These research results have led to various achievements in new technologies and applications. An object detection technique has been applied to the problem of face detection and has been integrated in a real-time prototype system. A lip-tracking algorithm has been developed and, in combination with motion recognition techniques, has been applied to visual speech recognition and visual person authentication. This approach has been combined with acoustic speech analysis methods leading to audio-visual speech recognition and audio-visual person authentication systems. Several methods in sensor fusion have been investigated and have been applied to multimodal person authentication systems. Our person authentication technology has been integrated by Cerberus AG and Ibermatica SA to potential application demonstrators. Other work in X-ray image analysis has addressed the extraction of articulators in X-ray image sequences. Finally, the group has also investigated several document analysis systems for hand printed, hand written, and cursively written text.

- **Machine Learning, including pattern classification, data analysis and knowledge extraction**.

  The main goal of this group is to maintain a strong expertise in advanced techniques that have been identified as being of direct interest to current and future work at IDIAP. This involves research in several areas as diverse as Bayesian learning, artificial neural networks, decision trees, hidden Markov models, support vector machines, and logical analysis of data.

  The fruitful cross-fertilisation between this wide base of learning techniques and specific applications in speech processing and pattern recognition has already led to some original approaches and interesting results. However, it requires an important research effort to adapt general methods to problems with characteristics such as large and noisy databases (speech databases).

  Finally, with the aim of identifying new promising research directions for IDIAP, the expertise of the group is also exploited in other, more prospective, activities such as time series prediction (as applied, e.g., to the prediction of financial markets or risks of avalanches), as well as to the design of assisted diagnosis systems.

- **System Management, including database management and prototype development**.

  The above three research groups are backed up by a System Management group, responsible for database management and prototype development, and working in close collaboration with the research groups in the case of more applied projects (such as national CTI projects).

As briefly described above, and as illustrated by Figure 3, the activities in the three research groups have been defined to be as complementary as possible, while fostering active collaboration across the different research themes.

While speech processing and computer vision are often complementary in (multimodal) applications, they are also often based on common theories and mathematical tools, and can benefit strongly from interaction. Some recent developments in handwriting recognition, for instance, have been using hidden Markov models, initially developed in speech recognition. Similarly, some recent advances in multi-stream processing (as pionneered by IDIAP, in collaboration with a few other laboratories like ICSI of Berkeley, USA, and FPMs of Mons, Belgium) will be exploited in both groups and will also directly benefit the development of multimodal systems (as recently shown at IDIAP by some preliminary work on audio-visual speech recognition).

The Machine Learning group provides additional theoretical support to the two other (more application oriented) groups by investigating new technologies that are common and useful to speech processing, computer vision, and multimodal processing. For example, in 1998 and 1999, the new methods proposed for decomposing a learning problem into sub-problems were successfully applied to the automatic speaker verification problem. Furthermore, research on the possible ways of combining

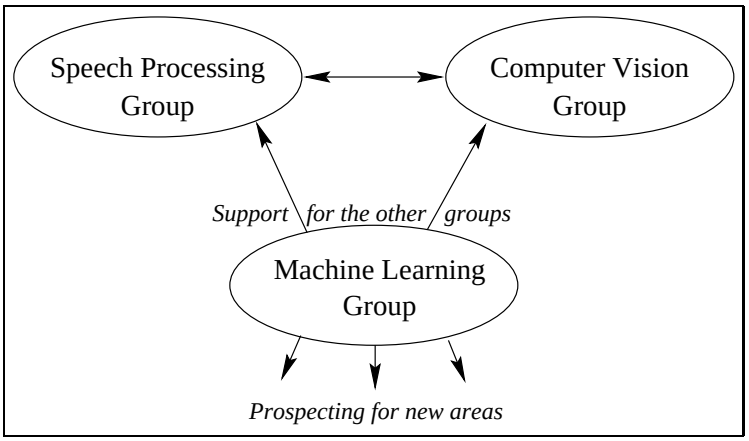



Figure 3: *Inter-dependencies between the three research groups.*

the sub-modules resulting from such a decomposition directly contributed to the problem of the fusion of multimodal experts. As a secondary goal, the Machine Learning group is also responsible for identifying and investigating new application areas which could directly benefit from the technology available at IDIAP (and primarily developed in the framework of speech and vision) and which could become important to future IDIAP activities (e.g., times series prediction). In this latter case, research will often be done in collaboration with other institutions with a larger expertise in the identified area.

## 1.3 Participation in National and European Community Research Projects

The activity of IDIAP within the framework of national (SNSF – Swiss national Science Foundation and CTI – Commission for Technology and Innovation) and European Community projects has been particularly intense for the last few years, and IDIAP has played a leading role in the conception and coordination of many projects.

While all these projects are briefly described in the present Activity Report (or see the introduction of the Activity Report of 1998 for of simple list of these projects), we give below the **list of new projects that were started or granted to IDIAP in 1999 only**:

- RESPITE: a European project, started on Jan. 1, 1999 for a three years period, aiming at the development of new techniques for noise robust speech recognition.
- InfoVOX: a CTI (Commission for Technology and Innovation), in collaboration with Swisscom, EPFL and VOXCom (spin-off of IDIAP), started on March 1, 1999, and aiming at developing advanced interactive vocal servers for tourist information.
- SV-UCP: a Swiss National Science Foundation (SNSF) project, started on February 1, 1999, and funding one PhD student working in the area of speaker verification based on user-customized password.
- INSPECT: an SNSF project, started on January 1, 1999, in collaboration with EPFL (Dr. Martin Rajman, EPFL/DI/LIA), funding one PhD student working on advanced integration approaches of acoustic and linguistic models in natural speech recognition.
- BN-ASR: an SNSF project, started on February 1, 1999, and funding one PhD student working on the development of new pattern recognition techniques (based on Bayesian Networks) applied to speech recognition.

- ARTIST-II: an SNSF project (follow-up of a previous project), started on April 1, 1999, and funding the last years of a PhD student working on the integration of articulatory features in state-of-the-art speech recognition systems.

- PROMO: an SNSF project, granted in September 1999 but not started yet, for the funding of one PhD student working on the modeling of pronunciation variants in speech recognition.

- SCRIPT: an SNSF project, granted in September 1999 and recently started, funding one PhD student working on very large vocabulary and subject independent cursive handwriting recognition.

- VOCR: an SNSF project, stared in October 1999, funding one PhD student on text recognition for video retrieval.

- ASSAVID (Automatic Segmentation and Semantic Annotation of Sports Videos), a European project (from the 5th RDT Framework), granted in August 1999 and starting early 2000, on the segmentation and indexing of sports videos.

- BANCA (Biometric Access Control for Networked and e-Commerce Applications), a European project (from the 5th RDT Framework), granted in August 1999 and starting early 2000, on the (biometric) user identity verification for internet applications.

- CARTANN: an SNSF project, started on January 1, 1999, in collaboration with the University of Lausanne, and funding one PhD student working on the modeling (and prediction) of geostatistic processes.

- DIVIDE & LEARN: follow-up of an SNSF project, funding the last year of a PhD student working on new pattern classification techniques based on the combination of multiple (neural network) models.

Most of these projects are discussed in more detail in the present report, at the end of the respective group description.

Finally, the involvement of IDIAP in a large SOCRATES/ERASMUS program (in which IDIAP was initially the only Swiss representative, later joined by EPFL) was renewed. The goal of this project is to define and initiate a European Masters program in language and speech. Common core courses will be given in all the represented European countries, followed by specialised courses and projects which will be given in specific institutions. In 1999, the content of Master programme was finalized and first implementations were initiated. At EPFL, this implementation should take the form of a postdoctoral course, which should start for the academic year 2000/2001.

## 1.4 Collaboration with other Organizations and Companies

Throughout the last few years, IDIAP has maintained close contacts with research organizations, universities, and industries working in the same research and developments areas. Those contacts typically originate from the follow-up of successful projects, or are based on personal long-term relationships and regular exchanges with some particular institutions. Just as a few examples, we can mention here:

- Strong partnerships with academic institutions such as EPFL (where Hervé Bourlard is also Professor), University of Geneva, and IMT (Univ. of Neuchatel). Several research projects involving a close collaboration between IDIAP and EPFL are currently going on, and several PhD students at IDIAP are, or will be, affiliated with EPFL. We are also initiating new projects with the University of Geneva.

- Initiated by European projects, IDIAP now has very good contacts with several companies, including Cerberus (CH), BBC (UK), Daimler-Benz (D), Thomson (F), Matra Notrel (F), Ibermatica (E) and several universities including Cambridge University (UK), Sheffield University (UK), Faculté Polytechnique of Mons (BE), University of Surrey (UK).

- Based on personal contacts and regular information exchange, we can mention here the active collaboration with Rutgers University (RUTCOR, USA) and the International Computer Science Institute (ICSI, Berkeley, USA), including student exchange.

## 1.5 Training Activities and Regional Development

On top of high quality research and development, we consider that two other major functions of IDIAP are:

**Training** and supervision of PhD students (most of the time affiliated with EPFL, University of Geneva, or University of Lausanne) and postdoctoral fellows, as well as short-term or medium-term visitors from academia (including ETS) and industry. As an example of this specific concern, IDIAP is currently working (initially as the only Swiss representative, but now joined by EPFL) with other European partners (in the framework of a European Socrates/Erasmus project) on defining the content of a European Masters in Language Technology (in which core courses would be given in all countries, followed by specialised courses and projects that would be taken in pre-defined countries). In 1999, the content of the Masters program was defined and first implementations have been initiated. At EPFL, and together with IDIAP, this resulted in the proposition of a post-doctoral course, which should first take place in the 2000-2001 academic year.

IDIAP is thus very active in the training of researchers and engineers, as well as in the training of highly qualified personnel in the scientific and technical fields. As of this writing, IDIAP is host to 13 PhD students, as well as several graduating students preparing their final thesis and coming from EPFL, Eurecom (F), ENST (F) and the ETS (Superior Technical School) of Sion. Every year, IDIAP also has a budget for 36 months (12 months for each group) of short-term visits for external fellows or students.

**Technology transfer** and industrial support, with two motivations: (1) to allow industries to keep up-to-date with the technology (since it has often become too expensive for even the largest companies to maintain in-house competence in all important areas), and (2) to allow IDIAP to develop and test prototype systems oriented towards applications of special interest to some of our sponsors. In this framework, IDIAP should also be able to perform qualitative and quantitative analyses, enhance the applicability of current base technologies, integrate technologies into pilot systems, and engage in active technology transfer.

Through open and intensive industrial collaboration, IDIAP aims to play an important role in promoting the economic development of the State of Valais. In this framework, VOXCom S.A. was started up in July 1998 as a direct spin-off of IDIAP with the collaboration of the City of Martigny. The primary emphasis for VOXCom is the development and deployment of software products and services designed to meet the specialized needs of businesses requiring Computer Telephony (CT), Call Centre functions, and integration support. VOXCom's emphasis is primarily integrating and leveraging base technology developed at IDIAP into turnkey solutions for information systems and other operational tools within the Call Centre environment.

Finally, in June 1999, IDIAP organized a one day symposium at "Hotel du Parc" in Martigny, to present its works, together with life demonstrations, and a few invited speakers working in related areas.

## 1.6 Publications

The quality of a research institution is also measured in terms of the volume of its (high quality) publications. Although this could still be improved, the number of publications in 1998/1999 has also increased compared to the years 1997/1998. For 1998/1999, these publications (listed in detail at the end of the present report) were the following:

- 7 books or book chapters
- 6 articles in international journals
- 61 articles in international conferences
- 33 internal research reports

# 2 Staff

**Mail:** IDIAP — Institut Dalle Molle d'Intelligence Artificielle Perceptive
CP 592
CH–1920 Martigny (VS)
Switzerland

**Phone:** +41 – 27 – 721 77 11

**Fax:** +41 – 27 – 721 77 12

**Internet:** http://www.idiap.ch

## 2.1 Scientific Staff

Persons at IDIAP in 1999 or as of this writing:

| | | |
|---|---|---|
| Dr. | Samy Bengio<br>Samy.Bengio@idiap.ch | Machine Learning group leader<br>+41 – 27 – 721 77 39 |
| Dr. | Souheil Ben Yacoub<br>Souheil.Ben-Yacoub@idiap.ch | research scientist<br>until November 1999 |
| Ms. | Giulia Bernardis<br>Giulia.Bernardis@idiap.ch | research assistant<br>+41 – 27 – 721 77 36 |
| Mr. | Olivier Bornet<br>Olivier.Bornet@idiap.ch | System Management group leader<br>until October 1999 |
| Prof. | Hervé Bourlard<br>Herve.Bourlard@idiap.ch | Director, Professor EPFL<br>+41 – 27 – 721 77 20 |
| Mr. | Datong Chen<br>Datong.Chen@idiap.ch | research assistant<br>+41 – 27 – 721 77 56 |
| Mr. | Thierry Collado<br>Thierry.Collado@idiap.ch | development engineer<br>+41 – 27 – 721 77 42 |
| Mr. | Beat Fasel<br>Beat.Fasel@idiap.ch | research assistant<br>+41 – 27 – 721 77 23 |
| Mr. | Frank Formaz<br>Frank.Formaz@idiap.ch | System Management group leader<br>+41 – 27 – 721 77 28 |
| Mr. | Dominique Genoud<br>Dominique.Genoud@idiap.ch | research assistant<br>until January 1999 |
| Mr. | Nicolas Gilardi<br>Nicolas.Gilardi@idiap.ch | research assistant<br>+41 – 27 – 721 77 47 |
| Mr. | Hervé Glotin<br>Herve.Glotin@idiap.ch | research assistant<br>+41 – 27 – 721 77 33 |
| Mr. | Frédéric Gobry<br>Frederic.Gobry@idiap.ch | research assistant<br>until September 1999 |

| | | |
|---|---|---|
| Ms. | Astrid Hagen<br>Astrid.Hagen@idiap.ch | research assistant<br>+41 – 27 – 721 77 34 |
| Prof. | Mikhael Kanevski<br>Mikhael.Kanevski@idiap.ch | research scientist<br>+41 – 27 – 721 77 49 |
| Mr. | Christopher Kermorvant<br>Christopher.Kermorvant@idiap.ch | research assistant<br>until November 1999 |
| Mr. | Sacha Krstulović<br>Sacha.Krstulovic@idiap.ch | research assistant<br>+41 – 27 – 721 77 43 |
| Dr. | Mikko Kurimo<br>Mikko.Kurimo@idiap.ch | research scientist<br>+41 – 27 – 721 77 41 |
| Mr. | Bertrand Liardon<br>Bertrand.Liardon@idiap.ch | development engineer<br>+41 – 27 – 721 77 48 |
| Dr. | Jürgen Lüttin<br>Juergen.Luettin@idiap.ch | Computer Vision group leader<br>+41 – 27 – 721 77 27 |
| Dr. | Djamila Mahmoudi<br>Djamila.Mahmoudi@idiap.ch | research scientist<br>until March 1999 |
| Mr. | Johnny Mariéthoz<br>Johnny.Mariethoz@idiap.ch | development engineer<br>+41 – 27 – 721 77 44 |
| Dr. | Eddy Mayoraz<br>Eddy.Mayoraz@idiap.ch | Machine Learning group leader<br>until August 1999 |
| Mr. | Mathew Magimai Doss<br>mathew@idiap.ch | research assistant<br>+41 – 27 – 721 77 57 |
| Mr. | Perry Moerland<br>Perry.Moerland@idiap.ch | research assistant<br>+41 – 27 – 721 77 32 |
| Dr. | Chafic Mokbel<br>Chafic.Mokbel@idiap.ch | Speech Processing group leader<br>until September 1999 |
| Dr. | Houda Mokbel<br>Houda.Mokbel@idiap.ch | research scientist<br>until September 1999 |
| Mr. | Miguel Moreira<br>Miguel.Moreira@idiap.ch | research assistant<br>+41 – 27 – 721 77 45 |
| Dr. | Andrew Morris<br>Andrew.Morris@idiap.ch | research scientist<br>+41 – 27 – 721 77 35 |
| Mr. | Bojan Nedić<br>Bojan.Nedic@idiap.ch | research assistant<br>+41 – 27 – 721 77 25 |
| Mr. | Todd Stephenson<br>Todd.Stephenson@idiap.ch | research assistant<br>+41 – 27 – 721 77 52 |
| Dr. | Georg Thimm<br>Georg.Thimm@idiap.ch | research scientist<br>until January 1999 |

Mr. Alessandro Vinciarelli — research assistant
Alessandro.Vinciarelli@idiap.ch — +41 – 27 – 721 77 24

Ms. Katrin Weber — research assistant
Katrin.Weber@idiap.ch — +41 – 27 – 721 77 37

## 2.2 Visitors

Mr. Jonathan Barker — University of Sheffield, UK
Jonathan.Barker@idiap.ch

Ms. Heidi Christensen — University of Aalborg, Denmark
Heidi.Christensen@idiap.ch

Mr. Raphaël Dupont — Saxon, Switzerland
Raphael.Dupont@idiap.ch

Mr. Manish Goyal — Indian Institute of Technology, Kampur, India
Manish.Goyal@idiap.ch

Prof. Timo Honkela — Media Lab, University of Art and Design Helsinki, Finland

Mr. Bruno Jakob — IRISA/INRIA, Rennes, France
Bruno.Jakob@idiap.ch

Mr. Sebastian Möller — Ruhr-Universität Bochum, Allemagne
Sebastian.Moeller@idiap.ch

Mr. Vesa Siivola — University of Technology, Helsinki, Finland
Vesa.Siivola@idiap.ch

## 2.3 Students

Mr. Eric Grand
Eric.Grand@idiap.ch from May 1999 to January 2000

Mr. Frederic Bressoud
Frederic.Bressoud@idiap.ch from October 1999 to January 2000

## 2.4 Administrative Staff

Ms. Sylvie Millius — secretary
Sylvie.Millius@idiap.ch — +41 – 27 – 721 77 21

Ms. Nadine Rousseau — secretary
Nadine.Rousseau@idiap.ch — +41 – 27 – 721 77 22

# 3 Research Activities

## 3.1 Speech Processing Group

**Group Leader: Hervé Bourlard and Chafic Mokbel**

The IDIAP Speech Processing group focuses its expertise on research and development in the area of speech and speaker recognition. This includes advanced research activities, maintenance of language resources for the training and testing of recognition systems, and development of real-time prototypes. The group has been involved in speech research for several years and is today at the leading edge of technology. As will be described in the following, it is involved in several national and European collaborative projects, as well as industrial projects.

### 3.1.1 Overview of the Speech Processing group activities

As illustrated in Figure 4, the activities of the Speech Processing group expand along two main axes: on one hand, the different research themes and underlying technology, and on the other hand, the development levels ranging from fundamental research to prototype development. The goal of prototypes is to demonstrate the added value of technology to different services, while providing important feedback to research. Nevertheless, the main focus of the Speech Processing group is kept at the technological level.

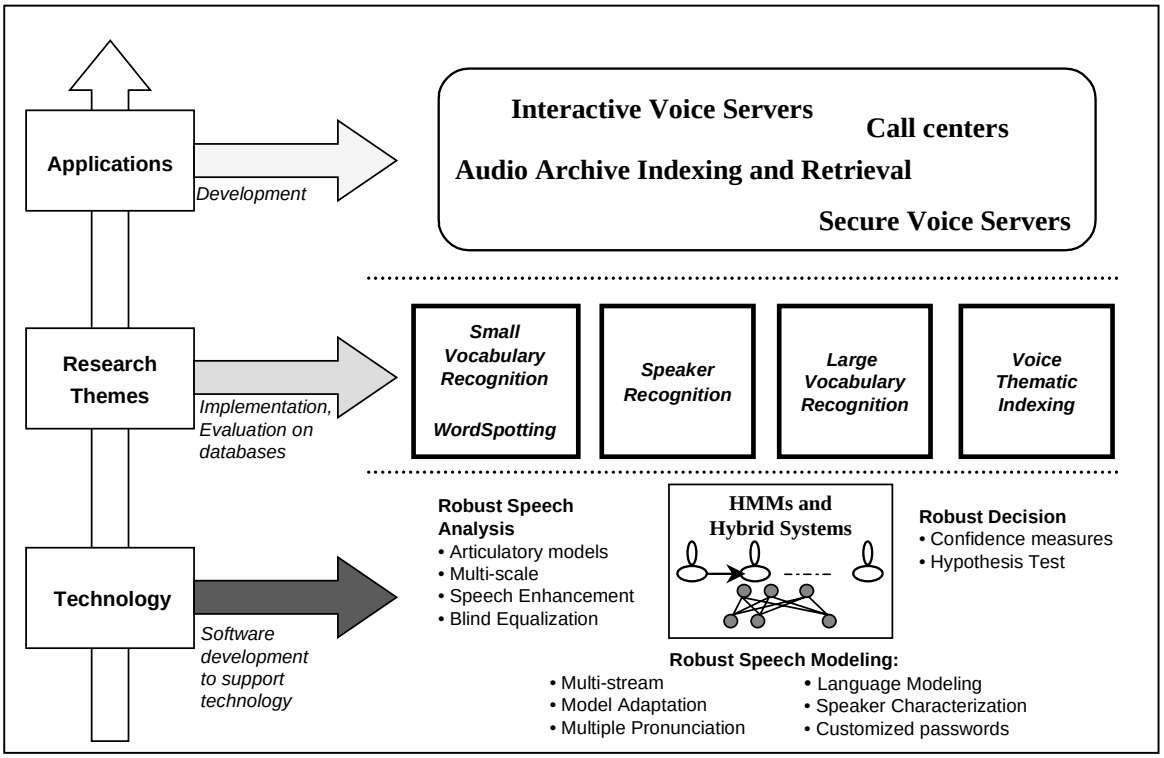



Figure 4: *Main activities within the speech group.*

**Technological level**

The technology of speech recognition is concerned with producing a word transcription that best matches an acoustic signal. In a few words, the technical core of the speech and speaker recognition systems studied at IDIAP is generally based on stochastic modeling. Typically, **Hidden Markov Models (HMM)** are used to model the distribution of the sequences of speech feature vectors obtained at the output of a speech analysis module. Standard HMMs, as well as variants such as **hybrid HMM/ANN** using Artificial Neural Networks (ANNs) together with HMMs, are particularly well mastered at IDIAP.

These HMM-based automatic speech recognition systems can achieve very high performance, and are now used in many real-life applications. Unfortunately, their performances remain very sensitive to the conditions of use: it will usually drop dramatically as soon as a mismatch arises between the acoustic or lexical or language model parameters, estimated on some training data, and the test conditions. Since it is impossible to forecast the test conditions at training time, different approaches are investigated at IDIAP to improve the overall performance of speech recognition systems. These approaches address the different levels of the speech recognition process, namely speech analysis (better acoustic features, more robust to noise), speech modeling (acoustic, lexical and linguistic modeling) and decision taking (e.g., hypothesis testing and confidence level).

As technology development strongly depends on the domain of application, IDIAP declines the generic technology into a wide range of speech recognition systems adapted to particular applications (e.g., with wider vocabularies, accounting for speaker dependence or speaking mode, performing speaker verification, etc.). In every application domain, specific speech and language databases are required to develop and evaluate the technology in the applicative context of interest. IDIAP carries out its developments on reference English or French databases, but also develops internal databases, such as Polyphone (recently collected for Swisscom).

**Research themes**

As shown in Figure 4, four main research themes, corresponding to the main application domains, have been defined and were investigated at IDIAP in 1999:

- **Small/medium Vocabulary Recognition**, which is generally used in simple voice command applications. In this case, research activities are mainly oriented towards better robustness to noise and channel distortions and remain at the acoustic analysis and acoustic modeling level. Activities here include articulatory modeling, multi-stream and multi-scale based speech recognition, speech enhancement, and missing data theory. Work is also done towards keyword spotting and rejection of out-of-vocabulary (OOV) words.

- **Speaker Recognition**, which consists in recognizing or verifying a speaker's identity from his/her voice. Here, we distinguish several approaches depending on the application constraints (e.g., text dependent/independent input, password or prompted text). IDIAP is at the leading edge in this field and is still pursuing advanced research in different directions, including decision module, customized passwords, new modeling strategies, and incremental enrollment.

- **Large Vocabulary Continuous Speech Recognition**. Work in this area addresses several modeling layers (acoustic, lexical and linguistic) as well as the interaction between these layers. At the acoustic level, better HMM and HMM/ANN models are investigated, taking advantage of some of the developments in small/medium vocabulary recognition. At the lexical level, better phonetic representation of the words, including context-dependent models and pronunciation variability modeling, are investigated. At the linguistic (syntactic) level, the use of high-order language models or new paradigms to interface the acoustic and linguistic modules is studied. Finally, the use of confidence levels is also investigated to rescore the N-best solutions at the output of the recognizer and to reject some unreliable words.

- **Voice Thematic Indexing**, which has been an important research focus at IDIAP in 1999.

Speech recognition can be used to automatically transcribe large audio databases. Information retrieval approaches can then be applied to the output of the recognizer to semantically index the database and facilitate its access (typically retrieving desired audio documents based on spoken or typed input queries). On top of our international collaboration, mainly concerned with English data (e.g., BBC broadcast news), we started developing a Swiss French recognition system and evaluated indexing/retrieving strategies in this context. Current research deals with information retrieval and discrimination between speech and music segments in audio documents. More specifically, Latent Semantic Analysis (LSA), generally used to build language models, is studied and adapted for indexing purposes. In view of performing recognition and indexing on speech segments only (and thus avoiding major sources of errors), several approaches are also being investigated for the detection of speech/music segments into audio files. This research was mainly performed in the framework of the European THISL project (see Section 3.1.9), and its use and interface are illustrated in Figure 5.

**Applications and development**

The last level of activities corresponds to the development of prototypes to demonstrate the added value of speech technologies through some examples of real-life services. Prototyping allows us to:

- Measure the real-life performance of the systems, as opposed to the performance observed on pre-recorded databases.
- Analyze the efficiency of the systems to pinpoint the aspects that require further improvement.

It therefore provides a very important feedback to research. The prototype systems are developed in close collaboration with the IDIAP System Management group. In 1999, and in line of the 1998 activities, three main applications were developed or improved: "Voice Dialing", "Personal Attendant", and "Voice Controlled Web Page Interface". In these systems, particular attention is paid to **Computer Telephony Integration** (CTI) i.e., the integration of computer and telephony features (e.g., forwarding voice mail to email).

In the following, we will start by briefly describing the basic speech recognition technology which is the basis for IDIAP's research, followed by a brief description of the current main research themes. Finally, Section 3.1.9 will present a short description of most of the ongoing projects in 1999.

### 3.1.2 Base Technology Tools

A typical speech recognition system is illustrated in Figure 6. At its input, the speech signal obtained at the microphone output represents the amplitude of the waveform as a function of time. After digitization (typically at 8kHz for telephone speech), this signal is analyzed to produce a sequence of feature vectors defining an information measure over time.

**Analysis module**

Spectral and homomorphic analysis techniques are most often used. Several signal processing algorithms can be used to perform this analysis, typically resulting in a feature vector every 10-ms. In practical state-of-the-art systems, further processing will be applied to the signal including, e.g., echo cancellation, channel effects reduction (reduction of convolutional and additive noise), speech enhancement, and begin-endpoint detection. At this analysis stage, additional transformations are applied to the features to reflect some speech perception and speech production properties.

**Acoustic modeling layer**

At the modeling stage, speech signal is characterized by a two-dimensional variability (temporal and spectral variability). Therefore, the acoustic realization of a word does not belong to a fixed dimensional space, and classical pattern classification algorithms cannot be used directly. Since no exact

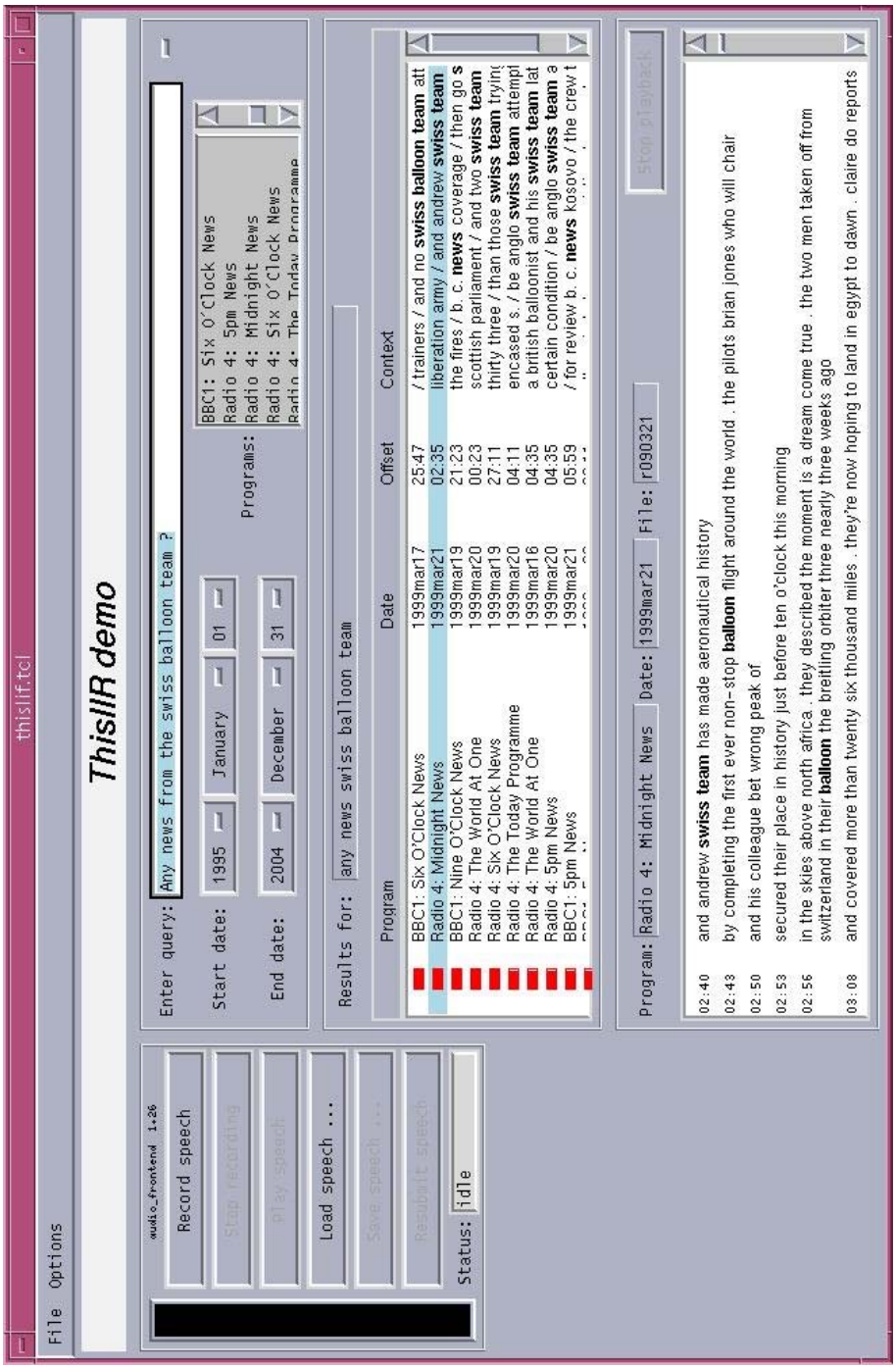



Figure 5: *Interface of the THISL audio indexing and retrieval system.*

physical model is available for the recognition of speech signal, stochastic models (sometimes referred to as "ignorance-based models") are generally used. In this case, it is assumed that speech sequences result from a Markovian process. Hence, Hidden Markov Models (HMMs) are popularly used to model speech units. Figure 7 represents a typical HMM, modeling a time-information distribution. In this

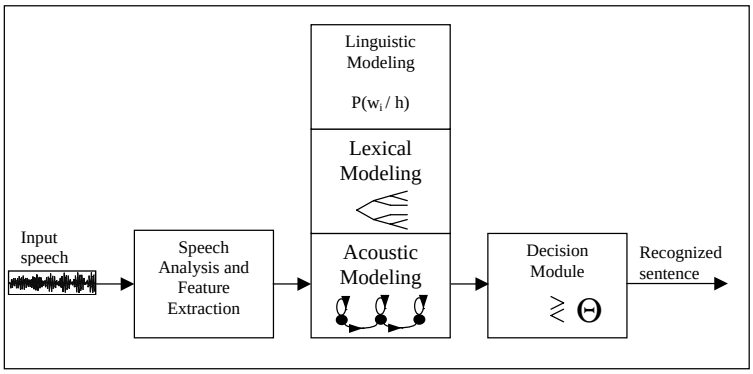



Figure 6: *Typical speech recognition system.*

framework, the speech signal is described as the output of a stochastic finite-state automaton built up from a discrete set of states. The system changes its internal state following a discrete probability density function known as transition probabilities. These are defined as the set of probabilities associated with the possible transitions of the underlying Markovian automaton. But the internal states of the underlying Markov process are not directly observable. Only the outputs of the process (the acoustic vectors) are observed and, given the frequency variability, these will be assumed to be themselves stochastic functions of the hidden internal states. Consequently, when in a specific state, the HMM system is assumed to emit measurable stationary observations according to a specific output distribution function. These output distributions are usually parameterized in terms of either Gaussian mixtures, or Artificial Neural Networks (ANN, in the case of hybrid HMM/ANN systems). The set of transition and output distribution parameters can be estimated through efficient training algorithms referred to as *Baum-Welch* and *Viterbi* training. These algorithms make use of large sets of training data constituted of acoustic sequences and their associated word sequences.

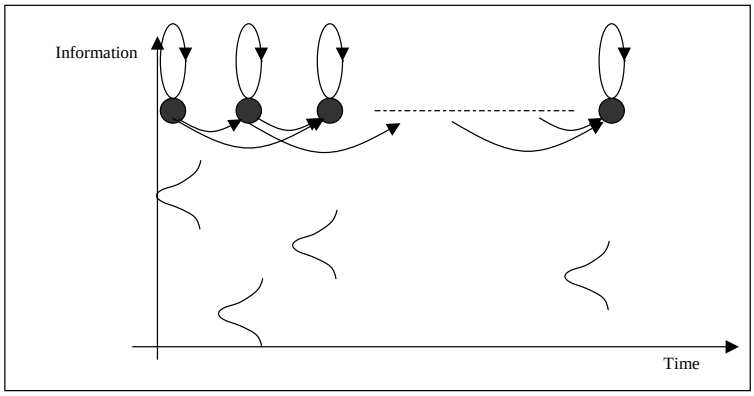



Figure 7: *HMM modeling of time-information distribution.*

**Lexical modeling layer**

HMMs can be used to model the words of a small vocabulary. But for large vocabularies, the number of parameters quickly becomes too excessive to allow for estimation on the basis of a reasonably

sized training set. Moreover, when the vocabulary size increases, the words become more confusable. These two reasons motivate the use of HMMs to model sub-word units rather than words, e.g., context-dependent phonemes or syllable models. In this case, the decomposition of words in terms of sub-word units adds a modeling layer to the system: the lexical layer. This increases the complexity of the decoder. For example, to consider pronunciation variants, several sub-word transcriptions can be used to define a word, and these transcriptions can be obtained from phonological rules and/or estimated from examples. Proper modeling of pronunciation variants is still an important research topic.

**Language modeling layers**

When going from isolated words to continuous speech recognition, additional modeling layers have to be integrated into the system. Their role is to take the grammatical constraints, and possibly semantic constraints, into account. This further increases the complexity of the modeling and decoding processes. A lot of work has already been done towards integrating different language models (LMs) into continuous speech recognition systems. This includes statistical LM (bi-grams and tri-grams), stochastic or deterministic finite state automata, as well as regular and context-free grammars (by using grammatical inference techniques). Most state-of-the-art systems today make use of statistical grammars, typically N-grams. These grammars model the probability distribution over all the lexicon words conditionally on a set of N previous words hypothesized during the decoding process. More complex syntactic or semantic constraints can be integrated by using the general framework of the N-best paradigm. In this case, the recognizer uses minimal syntactic constraints to generate a word lattice or a list of N-best sentences and then re-score (filter) the list by more complex linguistic models.

**Other layers and modules**

Large Vocabulary Continuous Speech Recognition systems require major adaptations of the HMM decoders (such as the Viterbi algorithm and the stack decoder) to:

- speed up the search (progressive search, beam search)
- allow for efficient lexicon/grammar representation, access, and integration.

In some cases, recognition systems will be complemented with an additional decision layer, ideally implementing the Bayesian decision rule. Indeed, while the recognition decision is pretty easy to take (acceptable complexity) in the case of close-set classification (typically close-set of isolated words), this complexity drastically increases when working on open-set classification problems. This is the case of continuous speech recognition with an infinite number of possible sentences, and/or when dealing with the problem of out-of-vocabulary (OOV) words rejection.

**Speaker recognition/verification**

Speaker recognition/verification systems make assumptions similar to speech recognition about the speech signal, and often use the same HMM-based techniques. It simply puts more focus on the decision stage and on the separation between the lexical content and the speaker characteristics in the speech signal. Thus, the base technology tools are the same as the ones used in speech recognition.

We have briefly recalled the architecture of a typical speech/speaker recognition system. This description clearly shows that speech recognition is a multi-disciplinary research field dealing with signal processing, information theory, stochastic modeling, speech production, speech perception, phonetics, linguistics and decision theory. IDIAP possesses a significant and an ever increasing competence in all these fields, and uses this competence to lead research for the enhancement of speech technology. IDIAP's research themes will be detailed in the following.

### 3.1.3 Small / Medium Vocabulary Robust Speech Recognition

In real-life applications, the speech signal presented at the input of a recognition system is often degraded with additive noise and/or channel distortions (convolutional noise). Moreover, the user

may pronounce words that do not belong to the target lexicon, especially in the case of small/medium size lexicons. Robustness to noise and out-of-vocabulary (OOV) words is a key factor to the success of automatic speech recognition in real-life applications. Considering Figure 6, robustness can be improved by acting on several modules of the system. Consequently, the main research directions at IDIAP in 1999 towards robust small/medium vocabulary speech recognition can be classified as follows:

- **Defining robust features :** Some acoustic features are known to be more robust to noise and channel distortions. Based on the autocorrelation function, short-modified coherence (SMC) is a representation robust to additive noise. Perceptual linear prediction (PLP) coefficients with RASTA filtering are resistant to channel distortions. Recently *J*-RASTA filtering was proposed to increase robustness to both additive noise and channel distortions.

  IDIAP has a large expertise with most of these features. In 1999, we assessed the robustness of RASTA-PLP and *J*-RASTA-PLP features to several types of additive noise and at different signal to noise ratios (SNR) on telephone databases. Another feature set was derived from articulatory modeling and was experimented : we developed an algorithm to perform acoustic to articulatory inversion in the case of the DRM model. Another approach that we also currently investigate is the use of multi-resolution features (defining multiple time scale features). These features are then combined directly into a single feature vector for further processing or integrated in the framework of the multi-stream approach described thereafter.

- **Developing preprocessing techniques** to reduce the effects of disturbances in the observed speech signals. Speech enhancement or blind channel effects equalization are used for this purpose. Spectral subtraction, based on the estimation of the noise spectrogram is also a very popular approach to reduce the effects of additive noise. However, since short-term estimation of signal-to-noise ratio (SNR) is not easy, this approach is best suited for stationary noise. In 1999, a variant of spectral subtraction has been developed at IDIAP in order to be combined with more recent developments based on multi-stream or missing data approaches.

  Cepstral mean subtraction (CMS) is very useful to reduce channel effects and was also experimented at IDIAP. In this framework, linear convoluted channel effects slowly varying with time are projected in the cepstral space as additive low-frequency components to speech cepstral trajectories.

  In 1999, IDIAP also started looking at blind equalization techniques using adaptive filtering.

- **Developing modeling architectures**, since robustness can be improved by using different **HMM topologies and models**. In 1999, IDIAP was intensively working on **multi-stream** and **missing data** approaches. The basic idea of the multi-stream approach is to represent the speech data in terms of independent subsets of feature vectors and to develop independent models for each subset. Since noise may affect a stream more than another, the robustness is increased by processing the different streams separately and recombining the output of the different models appropriately. A particular case of this approach is the multi-band technique, where the full frequency band is no longer considered as a single feature but as a set of sub-band features. Missing data theory is somehow related to this idea by assuming that noisy features are simply irrelevant. Hence, only the clean features, localized in time and frequency, are used in the calculation of the HMM probabilities. The two main issues underpinning these approaches are (1) how to identify the noisy or missing streams, or to measure their reliability, and (2) how to integrate this information into the decoding process. In the following, multi-stream and missing data approaches are further detailed, as well as the techniques initially developed to address these two issues.

  1. *Multi-Stream Approach*
     As exposed earlier, the acoustic processing module delivers a sequence of acoustic feature

vectors that each describes local components of the speech signal. HMMs then assume piecewise stationarity of the signal, and each stationary segment is associated with a specific HMM state. As illustrated in Figure 8, the multi-stream approach avoids this limitation by constructing a model for each subset of features, or feature stream. It thereafter combines the probabilities resulting from the different streams. This combination can be operated at different levels of the modeling process, e.g., HMM states, sub-word units (phonemes or syllables), words, or sentences.

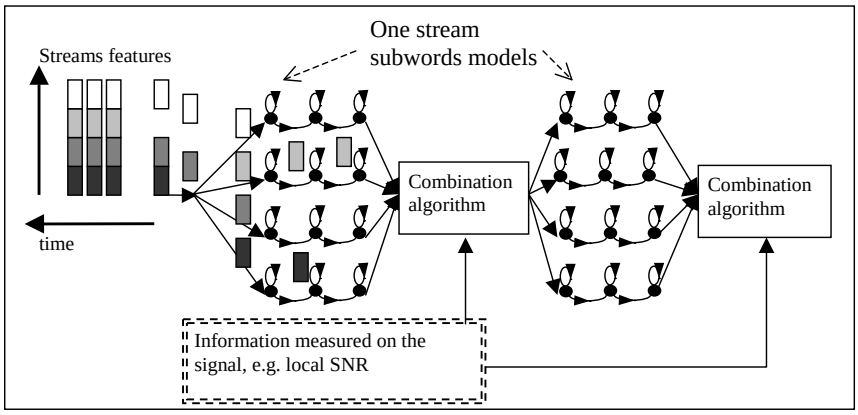



Figure 8: *Multi-Stream approach.*

2. *Missing Data Approach*
   The missing data approach is based on the assumption that some of the features are not observed, or are simply not relevant for recognition (too noisy or not carrying any information). For instance, in practice, features are never completely missing but highly disturbed by additive noise. In that case, the likelihoods associated with the current feature vector are estimated on the basis of the remaining features only (e.g., by estimating marginal distributions).
3. *Quantification of Noisy or Missing Streams*
   Localization and quantification of the noisy or missing streams remains a key issue to the success of both multi-stream and missing data, and several related approaches have been investigated at IDIAP in 1999. The first approach is based on sequential estimation of the noise characteristics, based on the estimation of first and second-order moments. To do this sequential estimation, the acoustic frames are classified as speech or noise through a statistical test. A second approach is based on the estimation of the local harmonicity degree. For voiced sounds, the harmonicity degree is defined as the ratio between the maximum R1 of the correlation function within a pitch period and the energy R0 of the signal.
4. *Integrating SNR Information into the Modeling Process*
   The estimated SNR must be integrated into the modeling process. This information will be used in the combination module of the multi-stream approach, or to estimate the set of likelihoods in the missing data approach. The first solution studied in 1999 consists in selecting the streams that are considered clean (not highly disturbed), or combining the likelihoods of the different streams according to a weighted sum, in which the weights are proportional to the estimated stream-specific SNR. Another solution, referred to as the *full combination approach*, performs the combination as a weighted sum over all possible stream subsets.
5. *Adaptive Speech Recognition Systems*

Adaptation of the model parameters to the actual condition of use is an important research direction towards better robustness of speech recognition systems. The main idea is to automatically estimate from the field data, and in an unsupervised way, new values for the model parameters to better match the statistical properties of the observed data. This approach will be investigated soon to complement the approaches discussed above.

All the preceding approaches have been studied and developed within the framework of several national and European projects. The relevant research grants are briefly discussed in Section 3.1.9.

### 3.1.4 Speaker Recognition

IDIAP has a strong interest in speaker recognition over the telephone network, with a particular focus on speaker verification. Speaker verification uses a customer's utterance to automatically verify a claimed identity. Since the system is based on stochastic modeling, it must be trained on each customer in an enrollment phase. During this phase, models will be built to allow for discrimination between the identity claimed by the user and possible impostors.

Speaker verification methods are divided into text-dependent and text-independent methods. The former requires the speaker to provide utterances of the keywords or sentences having the same text for both training and recognition trials. The latter does not rely on a specific text being spoken. In most cases, state-of-the-art speaker verification approaches are based on HMM techniques and greatly benefit from the progress made in speech recognition.

Although many recent advances and successes in speaker verification have been achieved, many problems remain to be solved. Most of these problems arise from variability, either originating from the speaker, or depending on channel and recording condition. From a human-interface point of view, it is important to consider how the users should be prompted, and how errors should be handled. Current speaker verification systems are not truly user-friendly, since they require long enrollment sessions. Moreover, for text-prompted systems, the choice of password is usually not flexible.

The research activities carried out at IDIAP in speaker verification can be described along several axes:

- **Improving the client/world modeling**. In order to verify the identity claimed by a speaker, two stochastic models are generally used : one for the claimed customer and the other for the "world" (where all the speakers different from the client). Given an utterance from a user, the likelihoods of both the client and the world models are computed and the ratio is compared to a predefined threshold. In order to improve the robustness of such a system, the client and the world models must be improved. In 1999, several algorithms were developed in that direction at IDIAP and were shown to improve the speaker verification performance :

  1. *Synchronous Speaker / World Alignment*
     Both client and non-client (world) hypotheses are modeled with HMMs. The decision is then based on the maximum likelihood between the two HMMs given the observed utterance. Separation of the HMMs assumes that the two models are independent, which is obviously not correct. A new structure is now being investigated where the speaker and the world are sharing the same HMM. We have established the theoretical foundation of such a model for optimal decoding and training. Figure 9 illustrates the principle of this new structure in comparison with the classical modeling, as investigated in the framework of the European PICASSO project (see Section 3.1.9).
  2. *Incremental Enrollment*
     HMMs adaptation algorithms can be used in an incremental enrollment strategy. As for speech recognition, adaptation allows to incrementally adapt the parameters of the models to better describe the data characteristics in the conditions of use. In the case of speaker recognition, these algorithms also allow to enrich the stochastic models initially trained on few enrollment utterances. These adaptation algorithms will also keep track of the inherent

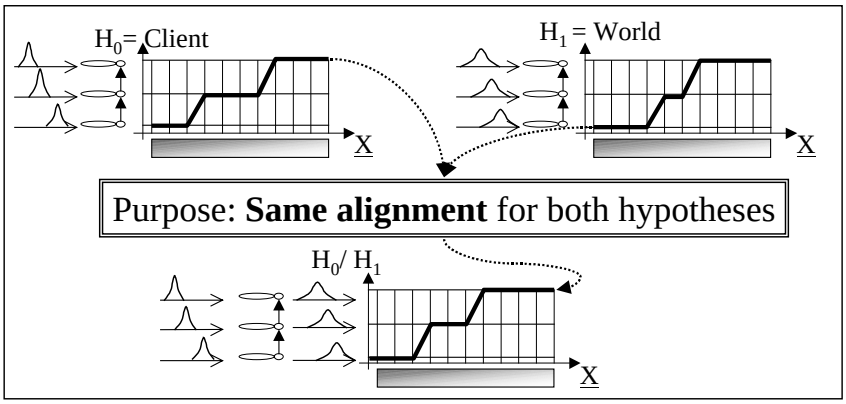



Figure 9: *Synchronous alignment approach.*

fluctuations of the customers' voices. Existing algorithms were studied and an API was defined for the implementation of these algorithms in the case of speaker verification.

3. *Customized Password*
   Text dependent speaker verification generally makes use of a password predefined by the system or the user. However, an added value can be obtained if the clients have the possibility to change their passwords easily. This requires the system to be able to infer the model of customized password. Several algorithms can be investigated in this context. These algorithms are similar to the ones developed for improving lexical modeling. An SNSF project (SV-UCP, see Section 3.1.9 is focusing on this project.

- **Improving the decision strategy**. Several issues may be studied to increase the robustness of the decision module. One important issue is the automatic selection of relevant acoustic vectors that yield robust discrimination between the customer and the world. Along this line, one approach has been investigated in 1999 regarding deliberate imposture. It consisted in automatically deriving imposture utterances that largely resemble the client utterances. Deriving such utterances can help adjusting the decision threshold and measuring the limitation of the state of the art speaker verification systems. A simple approach based on the concatenation of the client speech segments was investigated, and experimental results showed that current systems poorly resist such an imposture strategy.

- **Fusion of different systems**. Experimental results showed that the multiple speaker verification approaches available at IDIAP, based on different technologies, result in different recognition errors. Consequently, a possible solution to improve robustness of the systems consists in combining the scores of the different approaches. Along this line, several fusion algorithms were studied and experimented to merge the scores of the different systems. In the framework of the ELISA consortium, IDIAP participated in the 1999 NIST evaluation, and the fusion scheme resulted in leading edge performance.

All the preceding approaches have been studied and developed within the framework of several projects, which are briefly described in Section 3.1.9.

### 3.1.5 Large Vocabulary Robust Speech Recognition

As compared to small/medium vocabulary speech recognition, Large Vocabulary Speech Recognition (LVCSR) requires an improvement of robustness at both lexical and linguistic levels, plus improvements in the interaction of all the layers. The models are more complex and have to be manipulated

with care. In 1999, important work has been done at IDIAP to adapt the English hybrid HMM/ANN systems to perform robust LVCSR in French. The following directions can be distinguished.

- **Finding the optimal number of parameters to describe the acoustic distribution.** This number should bring a compromise between:
  - The precision of the resulting models
  - The reliability of the estimation
  - The complexity of the decoding process.

  In 1999, we have been focusing on hybrid systems. A large number of experiments have been conducted in order to determine the optimal size of the multilayer perceptrons (MLP). It was observed that significant improvements could be obtained from **enlarging the size of the MLP**. Furthermore, an alternative way of using the **training** of the MLP has been developed and provided satisfactory results.

- Introducing **pronunciation variants**. Multiple pronunciations for some of the vocabulary enriches the lexical modeling. The variants can be computed automatically for each word separately or using some rules and inference techniques. The rules are generally described using a tree or a network, which can be stochastic. Stochastic trees offer an elegant description of the rules augmented with probability values. Several approaches have been studied in 1999 and have shown that improving lexical modeling highly increases the global performance.

- **Increasing the reliability of language models.** Language models are usually trained from large text corpora. Generally, these text databases are not directly profitable to estimate the language models parameters. A **text preprocessor** has been developed to filter out the noisy sequences from the text data (e.g., headers, footers, transforming capital letters at the beginning of sentences to lower case letters, replacing "Mme." by "madame", ...). This significantly improved the quality of the estimated language models. More **precise language models** were also studied. For example, going from bi-grams to tri-grams or even to 4-grams increases the language modeling reliability, measured in terms of perplexity, as well as the recognition accuracy, measured in terms of error rates. However, increasing the complexity of language models generally increases exponentially the complexity of both decoding and training processes. A compromise has to be found. To go beyond the limitation of the classical N-grams models, which measures the probability of a word given its N-left-words, **long-span language models can be developed and/or a priori knowledge can be introduced**. The use of **multi-words** is an important step in this direction. Actually, some words that are often grouped together can be associated into a multi-word definition, included as a supplementary input to the lexicon. This is the case, for example, with the expression: "d'_autre_part". Multi-words can be chosen on linguistic criteria or on the basis of an automatic selection.

- **Interaction between the modeling layers**. Interaction between the acoustic, lexical and linguistic levels forms a main research topic. Generally, the computed a posteriori score is a function of the acoustic likelihood, the lexical transcription probability and the language probability. However, it is appropriate to weight differently these contributions given (1) the hypotheses made during the estimation of the local likelihoods, and (2) the lack of balance and the difference in reliability of the different models. To improve the layers' interactions, the use of an **acoustic/language-scaling factor** was investigated. This factor weights differently the acoustic and language components in the global score. A good choice of such factor leads to significant improvements, as shown by the experiments conducted in 1999.

- **Interaction with a Natural Language Processing (NLP) system**. In applications different from voice dictation, LVCSR systems are generally followed by a Natural Language Processing system that extracts the semantics from the pronounced sentence. In this domain, it is

preferable to provide the NLP module with a list of best solutions at the output of the recognition system, in the form of *N*-best lists or words lattices. The **NLP** module post-processes this list and **re-orders the solutions in order to get the most meaningful** one. To get better re-ordering, **confidence measures** must be associated with each part of the solution. Important work has been done at IDIAP in this direction. Relevant confidence measures have been developed and experimented with N-best decoding. They showed an increase of the global performance.

As a summary of the large LVCSR improvements achieved in 1999, we can mention the results obtained on the BREF database: starting from a WER of 33%, using the previously described approaches resulted in a 23% WER. This represents a **relative improvement of 30%**. These developments are now extented to the problem of natural speech recognition in the framework of the INSPECT project (Section 3.1.9).

### 3.1.6 Voice Thematic Indexing

Information retrieval (IR) of spoken documents decoded by a speech recognizer has a large field of application. To retrieve information from important audio databases, like broadcast news or touristic documentation, the documents must be transcribed and indexed. This is a costly task. Automatic speech recognition is very helpful in this respect. However, automating the task is a complex problem, since the vocabulary is very large and the speech is often spontaneous and disturbed by music.

Two main strategies permit to automate the indexing task. First, indexing the databases is possible by spotting the most informative words. This requires a system with a limited vocabulary and a high OOV rejection performance. Alternately, it is possible to recognize all the words present in the documents. This strategy requires a LVCSR system able to perform high accuracy recognition on spontaneous speech.

Besides the work on improving the robustness of LVCSR, which is described in the previous sections, two main directions are studied. The first one is related to the nature of the application, and concerns the need for an **information-retrieval algorithm** that is robust enough to recover the recognition errors. The second one is related to the nature of the original data, which is often mixed with noise and music. This opens the question of **speech/non-speech separation**, non-speech designing all the silence, noise and music segments of the signal.

Information retrieval algorithms measure the distance between a document and a particular request. Several decision criteria can be used: the simplest one is to find the requested word in the document. More elaborate stochastic retrieving algorithms compute some semantic distance between a document and the requested words. IDIAP has been working on this last class of algorithms. In 1999, IDIAP applied the basic **Latent Semantic Analysis (LSA) algorithm**, for retrieving purpose. LSA permits to infer a semantic cardinal continuous space where both important words and documents are represented as points. IDIAP extended the classical LSA algorithm using Self Organizing Maps (SOM) classifiers. One important problem, relative to retrieving information, is how to evaluate the different algorithms. IDIAP has proposed a new quantitative measure to **evaluate the retrieving algorithms**.

Speech / non-speech detection is generally done using adequate signal processing algorithms. Several approaches and several features have been proposed in the literature. IDIAP has started developing a **2-steps algorithm for this segmentation**. First the signal is segmented into silence/non-silence parts. Then each part is classified as speech or music. An effort has been dedicated to label a database collected for research purpose. All the preceding approaches have been studied and developed within the framework of the THISL project briefly described in Section 3.1.9

### 3.1.7 Prototyping and Spoken Language Resources

Prototyping and spoken language resources are necessary to support technology developments. Building prototypes permits to observe the technology in a real-life environment. By the same way, it allows

to understand its relative added value and weak points to improve. Finally, it permits to measure the acceptability of the technology by end users. As a matter of fact, no technology development and evaluation can be done without spoken language resources. For these reasons, IDIAP has a main interest into these two components of the technology development.

A Prototyping speech recognition application has been realized within the AVIS (SWISSCOM) project. Two main applications have been developed: **voice dialing** and **personal attendant**. These prototypes concern the call completion phase in a telephone cycle. They completely integrate speech recognition functionality within the application. These prototypes will be used to test the functionality of the different systems, and to study the impact of speech recognition on telephone services. By developing these prototypes, IDIAP increased its expertise in **Computer Telephony Integration (CTI)**. This concerned, for instance, the manipulation of the different facilities within the ISDN protocol. Independently of the SWISSCOM project, IDIAP has developed a system for voice navigation on the WEB pages. Another prototype, showing voice-indexing capabilities, has also been developed (see Section 3.1.6).

### 3.1.8 Software Development

As exposed above, IDIAP has participated to the development of several general-purpose software tools in collaboration with other laboratories, mainly in the framework of European projects. Towards the end of 1999, an important development activity has been initiated at IDIAP in order to write completely new IDIAP software, with the goal of supporting our general and specific research and development needs. A team involving researchers from different groups has been set up to develop and cross-validate the new software. Several modules have already been written. This development activity aims at creating a system that integrates both classical and hybrid HMM systems. It will handle small and large vocabulary recognition of isolated words and/or continuous speech. The new system will facilitate the integration of the new approaches developed at IDIAP.

### 3.1.9 Research Grants

♢ MULTICHAN – Non-stationary MULTI-CHANnel signal processing

**Funding:** Swiss National Science Foundation (SNSF)

**Duration:** January'98 – December'99

**Persons involved:** Katrin Weber, Andrew Morris, Hervé Bourlard, and Samy Bengio

**Description:** The purpose of this project is to investigate a new multi-channel signal processing technique, which has recently shown much promise in the framework of multi-band speech processing. In multi-band speech recognition, the frequency range is split into several bands, and information in the bands is used for phonetic probability estimation by independent modules. These probabilities are then combined for recognition later in the process, at some segmental level. This multi-band paradigm is motivated by psycho-acoustic studies and by its potential robustness to noise. However, research in this important new approach is still preliminary. The current project thus started by investigating important issues related to this approach, including trade-offs between segment choices, features, and recombination approaches.

Furthermore, the same multi-channel paradigm will also be used to address the problem of multiple time scale analysis (e.g., towards incorporating multiple time scale information) in current speech recognition systems. The multi-channel approach considered here is a pretty new research area. It is however already attracting a lot of interest and could have an important impact not only on speech recognition research but also on many problems dealing with complex non-stationary temporal signals.

**Achievements:** A reference speech recognition system based on the HTK software platform was developed. During the tests on the Numbers95 telephone database it proved to be one of the best systems. Furthermore, multiple time scales speech analysis was implemented. Different analysis were combined into feature vectors, which were again transformed to lower dimensional vectors. The resulting performance is very encouraging since this method showed to be robust in presence of noise. Work will be pursued in this direction to select the most informative features from the different scales.
In the case of sub-band speech recognition, a novel technique for speech recognition in noisy environments, referred to as the "Full Combination" method, was developed to avoid independent sub-band processing. Recognition results for speech in noise were better than for any previous sub-band ASR method. This method also has considerable scope for further improvement through improved methods for local data reliability estimation. Full combination approach has been developed within both MULTICHAN and SPHEAR projects.

◊ SPHEAR – SPeech, HEAring and Recognition

**Funding:** European project, DGXII, TMR Research Network, supported by OFES

**Duration:** March'98 –February'2002

**Persons involved:** Astrid Hagen and Christopher Kermorvant

**Description:** SPHEAR is a four years European TMR project involving several European laboratories: Sheffield University (UK), Daimler-Benz (Germany), Ruhr-Universität Bochum (Germany), Institut National Polytechnique de Grenoble (France), University of Keele (UK), University of Patras (Greece), and IDIAP. The twin goals of this research network are to achieve better understanding of auditory processing and to deploy this understanding in automatic speech recognition in adverse conditions. This project has several themes, including computational auditory scene analysis, sound-source segregation and new recognition techniques based on multi-band and multi-stream processing.

**Achievements:** The basic multi-stream processing and the missing data approach have been implemented within two different speech recognition systems (HTK and STRUT). The first system (HTK) is based on standard HMM modeling with Gaussian mixture as output distributions while the second system (STRUT) implements a hybrid HMM/ANN system in which the output distributions are approximated by Artificial Neural Networks (ANN, in our case a Multilayer Perceptron, or "MLP"). The basic versions of multi-stream and missing data have shown to achieve competitive performance on Numbers95 disturbed by additive noise extracted from the NOISEX database. Important effort has also been put into the definition of an experimental protocol allowing a fair comparison of the different algorithms. Improved algorithms for feature reliability estimation and integration were proposed, as described in the introduction of this section, resulting in significant improvements with respect to the basic versions of the algorithms. Spectral subtraction and *J*-RASTA-PLP were evaluated as reference systems. Figure 10 reflects some of the results obtained in 1999.

◊ ARTIST-II – Articulatory Representations To Improve Speech Technologies

**Funding:** Swiss National Science Foundation

**Duration:** April'99 – March'01

**Persons involved:** Sacha Krstulovic

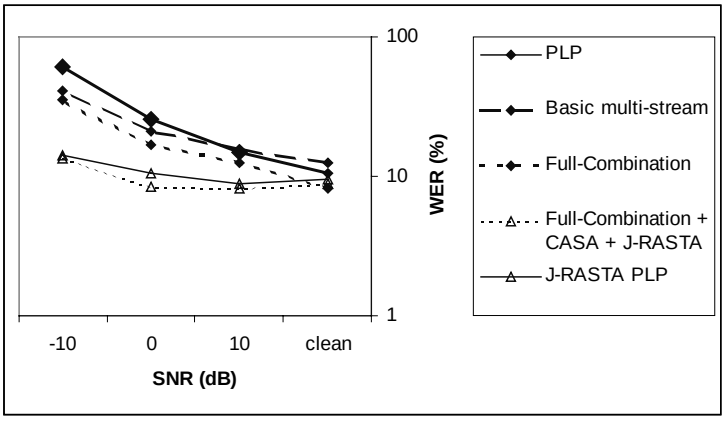



Figure 10: *Some results obtained in 1999 for robust speech recognition.*

**Description:** This project involves both speech and vision groups. For the description of the project please refer to 3.2.7.

**Achievements:** In order to use articulatory representations to improve speech recognition, relevant articulatory parameters should be automatically extracted from the speech signal. This is an acoustic to articulatory inversion problem. This inverse filtering problem is particularly complex since:

- It is largely non-linear
- It does not yield a unique solution if articulatory context is not taken into account.

In the framework of the present project, we derived an optimal solution to this problem in the case of the Distinctive Regions and Modes (DRM) articulatory model. This solution is based on lattice filtering and inverse filtering and introduces constraints on the classical autoregressive model. This inversion solution achieves the introduction of a true articulatory speech production paradigm in several domains of speech processing such as speech analysis, speech coding, speech synthesis and, speech recognition. For the first time, articulation-based recognition experiments were conducted using this representation.

◊ PICASSO – PIoneering Caller Authentication for Secure Service Operation

**Funding:** European project, Telematics project from DGXIII, supported by OFES

**Duration:** March'1998 – February'2001

**Persons involved:** Johnny Mariéthoz, Bojan Nedic, Chafic Mokbel

**Description:** PICASSO builds upon the work done in the CAVE project, which has improved speaker verification technology and has performed security experiments with a range of prototype implementations. It is a 3 years project that involves the CAVE partners i.e., IDIAP, Ubilab (CH), Swisscom (CH), ENST (F), IRISA (F), PTT-Telecom (NL), KPN Research (NL), KUN (NL), Fortis (NL), KTH (SE), Telia (SE), and Vocalis (UK). Work within CAVE highlighted that a tradeoff between the provided level of security and the usability of the system has to be found. Voice based verification does not pose hardware device problem for users, since all that is required is a standard telephone and since the

number of telephones available is being swelled by the enormous growth in the mobile and GSM markets.

PICASSO aims at integrating verification with automatic speech recognition to develop a new generation of telephone enquiry systems, that would combine high-accuracy customer verification with easy-to-use speech recognition interfaces. Project's results will be applied to telephone calling cards/accounts, messaging services ("voice mail") and retail banking services.

**Achievements:** within the PICASSO project, IDIAP is mainly involved in the improvement of the state of the art speaker verification algorithms. In this context, IDIAP distributed the Polyvar database to the partners in 1999, participated actively in the setup of the reference system, and participated actively to the definition of the corresponding experimental protocol. IDIAP developed both the theoretical and software parts of the synchronous alignment algorithm described earlier. The results are shown in Figure 11. IDIAP defined the software API and algorithmic issues related to incremental enrollment.

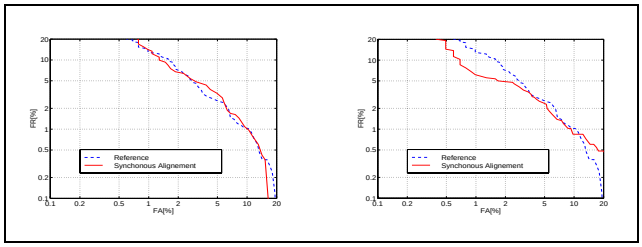



Figure 11: *Examples of results obtained with synchronous alignment.*

◊ COST249 – Automatic Speech Recognition over the Telephone

**Funding:** European project, COST action, supported by OFES

**Duration:** October'98 – September'00

**Persons involved:** Giulia Bernardis and Hervé Glotin

**Description:** The COST249 action involves several European laboratories and covers most of the European countries. This collaborative COST action aims at improving state-of-the-art speech recognition systems over the telephone network. It is a very broad project, addressing all-important aspects of continuous speech recognition systems, namely:

- Concept establishment: overall system configuration, task complexity and dialog modeling.
- Linguistic processing: lexical knowledge, parsing strategies, higher order constraints, language models, and speaker adaptation.
- Phonetic decoding: neural networks and HMMs, task and language independence, and recognition units.
- Acoustic signal processing: feature extraction, noise suppression, and speech corpora.

In the framework of this COST action, IDIAP is more particularly involved in:

- The development and improvement of acoustic decoding algorithms for continuous speech recognition over the telephone
- Their integration with higher level knowledge such as phonological and syntactical constraints.

At the national level, this work is carried out in collaboration with ETH.

**Achievements:** In 1999, IDIAP has largely improved its LVCSR performance. This was achieved by

- Optimizing the MLP size and MLP training
- Using multiple pronunciations to describe words in the lexicon
- Preprocessing the text databases used for training the language models
- Replacing the bigrams with trigrams as language models.

A new algorithm for the computation of the confidence measures has also been proposed.

♢ THISL : THematic Indexing of Spoken Language

**Funding:** European project, ESPRIT Program, Long Term Research supported by OFES

**Duration:** February'97 – January'2000

**Persons involved:** Mikko Kurimo

**Description:** THISL is a 3-years European project in which IDIAP is involved with different partners: Sheffield University (UK), Cambridge University (UK), Thomson (FR), BBC (UK) and ICSI (Berkeley, CA, USA). The objective of THISL is to show the feasibility of integrating state of the art Natural Language Processing (NLP) and Large Vocabulary Continuous Speech Recognition (LVCSR) technologies towards advanced multimedia applications. In this framework, the project focuses on R&D aimed at retrieving multimedia information (written or spoken text) using a spoken language interface. Most of the tests will be performed on recordings of BBC broadcast news.

**Achievements:** In 1999, IDIAP has largely improved its LVCSR performance for French recognition as shown in the Section 3.1.5. IDIAP has also adapted the THISL prototype to demonstrate the retrieving capacities on a French voice database. At the research level, IDIAP has developed several variants of the LSA algorithm for retrieving information and has shown the usefulness of the perplexity measure to evaluate and to compare the retrieving algorithms. IDIAP has studied several algorithms for the speech/music detection problem.

A real-time prototype system is now available for navigating in the sound-track of a TV news broadcast, with the following capabilities:

- Automatic recognition, transcription and indexing of BBC broadcast news
- Audio editing tools
- Content-based retrieval from audio archives (generated automatically)

♢ RESPITE – Recognition of Speech by Partial Information Techniques

**Funding:** European Project, DG XXIII

**Duration:** January'99 – December'2001

**Person involved:** Andrew Morris, Hervé Bourlard, Hervé Glotin

**Partners:** Sheffield University (UK), Daimler-Benz (D), MATRA (F), Polytechnic University of Mons (B), Univ. of Grenoble (F), and Intl. Comp. Science Institute (US).

**Description:** The main goals of this project are the following: Objectives:

1. To develop new techniques for automatic speech recognition that are truly robust to unanticipated noise and corruption, based on emergent theories of decision-making from multiple, incomplete evidence sources and of speech perception in listeners. More specifically, new recognition paradigms based on multi-stream processing and the missing data theory will be investigated here.
2. To test and deploy these techniques in two application areas: cellular phones and recognition in cars.

The expected results of this project are: (1) an extension of the range of conditions under which speech recognition can be used in general, and specifically of cellular phones and recognition in cars, and (2) advances in adjacent recent fields, such as the handling of multiple temporal resolutions and the processing of information multi-modal technology (e.g., audio-visual fusion).

**Achievements:** During the first year of the project, different baseline systems (including standard HMM, the AURORA reference HMM, and HMM/ANN systems) have been tested on common noisy databases, including connected digits (TIDIGITS), free format numbers and the car noise database from Daimler-Chrysler. On top of further developments of the multi-stream approach (as investigated in the SPHEAR project), a variety of local SNR estimation techniques, together with missing data theory, were tested. Furthermore, several other methods for estimating substream reliability (based on data likelihood or harmonicity measures) were investigated.

♢ InfoVOX – Interactive Voice Servers for Advanced Computer Telephony Applications

**Funding:** Swiss Commission for Technology and Innovation (CTI)

**Duration:** March'99 – February'2001

**Person involved:** Frank Formaz, Thierry Collado, Bertrand Liardon, Giulia Bernardis, and Olivier Bornet.

**Partners:** EPFL (DI/LIA), Swisscom, VOXCom S.A., and Omedia S.A.

**Description:** The objectives of this project are twofold:

1. Doing further research and development in the field of interactive voice servers, with applications in the key area of call centers for computer telephony applications.
2. Involve and support (through more R&D) a start-up company (VOXCom) developing computer telephony applications, and integrating complete call center solutions.

More specifically, the generic goal of this new project is to improve state-of-the-art automatic speech recognition and speaker verification technology (acquired in the framework of several national and international projects, as well as in the previous CTI project) and integrate this technology in a few, well defined, computer telephony application prototypes. The targeted application has been carefully designed by the partners to represent a good, balanced, mix of (re-usable) developments, including research, development of software tools (e.g., database access and user interfaces), integration, application, and application testing. More specifically, the generic application will be the development of Interactive Voice Response (IVR) systems (interactive vocal query systems) to access large and complex (possibly distributed) information databases. In the current project, we will mainly focus on internet databases, and as a testbed we will develop an voice interface to the web page available from the tourist bureau of Martigny (http://www.martigny.ch).

**Achievements:** In 1999, IDIAP mainly worked on:

1. The training and development of a baseline recognizer in French adapted to the InfoVOX task, including the acoustic models (trained on the Polyphone database) and the language model (initially trained on transcribed "Wizard-of-Oz" data).

2. The development of the telephone interface based on a Dialogic board.
3. More research on confidence levels.

At the end of the first year, an initial prototype system able to recognize spoken queries about restaurants in Martigny is available. In this (non realtime) prototype, spoken queries pronounced through the PC-based Dialog interface can be sent to a SunStation performing recognition and returning the result of recognition to the PC for further processing and interface with the user. A first version of the dialog model (complemented by a context-free grammar parser) has been developed by EPFL and will be soon interfaced with the recognizer.

♢ INSPECT: INtegrating SPeech (acoustic and linguistic) ConsTraints for enhanced recognition systems

**Funding:** Swiss National Science Foundation

**Duration:** January'99 – December'2000

**Person involved:** Giulia Bernardis, Hervé Bourlard, and Martin Rajman (EPFL/DI/LIA)

**Partners:** EPFL (DI/LIA)

**Description:** The main research goal of the present project is to develop and assess new strategies for integrating state-of-the-art acoustic models and advanced language models (LM) into speech understanding systems, in view of improving dialog-based interactive voice response (IVR) systems.

Interfaces between continuous speech recognition systems and advanced language models typically use the $N$-best paradigm, based on the maximum likelihood criterion. Unfortunately, the resulting hypotheses do not necessarily contain much semantic variability, and are not well suited to dialog-based systems. Consequently, the general research theme of the current project is to investigate new ways of generating $N$-best hypotheses that include more "semantic" variability, becoming therefore more appropriate for linguistic post-processing.

The applicative framework that will be used for the evaluation of the performance will be the one currently developed in collaboration between EPFL, IDIAP, ISSCO and Swisscom for the design and implementation of a dialogue-based information system for Advanced Vocal Information Services.

**Achievements:** During the first year of this project, we mainly focused our work on the following tasks:

- Acoustic modeling of telephone-based Swiss-French speech.
- Lexical modeling.
- Linguistic modeling of spontaneous requests.
- Improvement of interaction between acoustic, lexical and linguistic modeling layers.
- Setting up of a reference system for **Large Vocabulary Continuous Speech Recognition (LVCSR)**, with initial tests on the 111 service calls subset of the Swiss-French Polyphone database.
- Preliminary experiments with different recognizers (using different information) to be considered as a mixture of experts.
- Development of efficient modules for assigning confidence scores to hypothesized transcriptions and rejection of uncertain data.

♢ SV-UCP: Speaker Verification based on User-Customized Password

**Funding:** Swiss National Science Foundation

**Duration:** January'99 – December'2000

**Person involved:** Bojan Nedic, Hervé Bourlard

**Description:** The general objectives of the present project is to further improve state-of-the-art speaker verification systems in which IDIAP has shown to be at the leading edge.

More specifically, starting from the good background acquired by IDIAP in the framework of the previous FNRS project on speaker verification (soon resulting in a PhD thesis), the aim of the current project is to investigate new alternatives to speaker verification systems, including: (1) user-customized speaker verification systems (allowing the user to choose his/her password, just by pronouncing it a few times), and (2) automatic adaptation of the system. For many reasons that will be discussed later, this research will be carried out in the framework of a new technology merging HMM with artificial neural networks (ANN), and which has already shown much promise in the case of speech recognition.

This work should take place in the framework of several industrial and European projects which should highly benefit from this more fundamental research project.

**Achievements:** After initial works on text-independent speaker verification systems (including the NIST competition), an HMM inference software was developed and adapted from PICASSO to allow the inference of word HMM models (user-customized models) in terms of peudo-phonetic strings based on one or two pronunications of the user-specific password. A new neural network (MLP) training program was also written and adapted to allow speaker adaptation of the parameters. Starting from a speaker independent network (trained on a large speaker independent database), each speaker-specific neural network can be adapted on the enrolment sentences containing the user-customized password to maimize the likelihood of the infered HMM models. Initial tests to measure the discriminant capabilities of these networks with respect to (1) the same speaker (customer), (2) the world model (speaker-independent network), and (3) impostors were carried out.

♢ BN-ASR: Modelling the hidden dynamic structure of speech production in a unified framework for robust automatic speech recognition

**Funding:** Swiss National Science Foundation

**Duration:** March'99 – February'2001

**Person involved:** Todd Stephenson, Andrew Morris, Hervé Bourlard, and Samy bengio

**Description:** The main goal of this project is to develop new acoustic/phonetic models for speech recognition. Although Hidden Markov Models (HMMs) has been, for years, the most successful speech recognition approaches, these are rather simplistic stochastic models that only vaguely reflect the nature of speech. This project will extend the hidden state space of HMMs in various ways to better represent the hidden structure of speech production.

Bayesian Networks, relatively unknown in ASR, will serve as a framework for dynamic stochastic modeling. Thus the project will benefit from past and current developments for Bayesian Networks, and is expected to contribute to this area as well.

The project will interact with other projects at IDIAP concerning the influence on speech production caused by prosody, speaker characteristics, and articulatory constraints. These information sources will be incorporated in the stochastic model in addition to the usual phonetic information.

**Achievements:** An extensive literature review has been done, which was especially important as Bayesian networks is a new field to IDIAP. Upon gaining an understanding of the field, a report (which is almost complete) summarizing Bayesian network theory has been prepared.

Extensive software development has also be carried out to allow training and testing (inference of probability values of missing variables given observed variables) of generic Bayesian networks. This software is currently being extended to dynamic Bayesian networks to allow training and testing on speech recognition problems. After initial tests, and comparisons with state-of-the-art approaches, this approach will be extended to other tasks to take advantage of the properties of Bayesian networks (e.g., better modeling of probability density functions and correlation between variables, and explicit use of a priori constraints). For example, articulatory parameters could be included as additional variables to the speech recognition system.

♢ PROMO – PROnunciation MOdeling for automatic speech recognition

**Funding:** Swiss National Science Foundation

**Duration:** 2 years

**Person involved:** Nobody yet

**Description:** The generic goal of the present project is to develop new approaches towards pronunciation modeling in modern speech recognition systems. Most of these ASR systems model words as sequences of subword units, typically phoneme models, represented in a dictionary containing one or a few pronunication variants.

However, our recent experience with natural telephone speech (Polyphone) and large vocabulary microphone continuous speech recognition emphasized the weakness of the current pronunciation modeling approaches. The goal of the present project is thus to allow one PhD student to work full-time on this aspect of speech recognition and to investigate novel pronunciation models that will carefully consider the tradeoffs between expressiveness and confusability. Consequently, on top of further investigating standard approaches, this project will focus on (1) *dynamic* pronunciation modeling, and (2) discriminant training of pronunciation models.

♢ SOCRATES – European Masters in Language and Speech

**Funding:** European Project, DG XXII

**Duration:** September 97 – September 2000

**Person involved:** Hervé Bourlard

**Description:** The purpose of this project is to organize an advanced course (recognized as a European Masters degree) allowing students to qualify for multidisciplinary team-working in the language industries. The project involves Univ. of Saarlandes (D), Aalborg Univ (DK), Univ. of Sheffield (UK), Univ. of Essex (UK), Univ. of Edimburgh (UK), Univ. of Brighton (UK), Univ. of Athens (GR), Univ. of Patras (GR), Univ. of Nijmegen (NL), Univ. of Utrecht (NL), Univ. of Lisbon (P), IDIAP-IKB (CH) and EPFL (CH). Besides in depth knowledge of Speech Science, Natural Language Processing or Computer Science, provided by undergraduate studies, the student will obtain contextual knowledge from the fields that were not part of his/her specialization. IDIAP has the objective to create a center of excellence in the domain of Speech Processing for graduated students. This center is expected to become part of a large European teaching network. At the European level, this would cover well-defined common courses, taught in every participating country, as well as specialized courses and research projects, in countries where special expertise has been identified.

## 3.2 Computer Vision Group

**Group Leader: Juergen Luettin**

**Basic Research** The Computer Vision group studies problems in computer vision including object detection and recognition, motion analysis and recognition, shape analysis and representation, sensor fusion, and document recognition.

**Technologies** Recent research activities have mainly focused on topics in multimodal signal processing which is largely the results of the complementary expertise of both the members within the group and across groups at IDIAP. We are working on topics such as multimodal speech recognition, multimodal biometrics, handwriting recognition, X-ray image analysis, face detection and tracking, and facial expression recognition. Several of these topics have been investigated in collaboration with the speech processing and machine learning groups, taking advantage of different disciplines and complementary expertise.

**Applications** Our research topics are usually motivated by particular applications, mainly in the area of multimodal interfaces, biometrics, and multimedia document processing. Some projects are conducted in collaboration with industries and academic partners, usually in the framework of European projects, others are financially supported by the Swiss National Science Foundation (SNSF). During 1999, the group has been granted two new SNSF projects and 2 new EC projects.

### 3.2.1 Multimodal Biometrics

With the steady increase in applications and services that require secured access control, reliable personal identity verification is becoming more and more important. Traditional person authentication system based on personal identification numbers (PIN) or identity cards (ID cards) often don't meet the high requirements of security applications as (i) they can easily be transferred to other persons and (ii) the authentication system is not able to differentiate between and authorised person and an impostor that fraudulently uses a PIN or ID. Biometrics has the potential of overcoming these shortcomings by the use of biometric information that is proper to an individual which therefore can not easily be transferred to another person. Biometric technologies that obtain wide user acceptance are, as example, based on face, voice, signature or thermogram. However, these methods lead only to limited authentication performance.

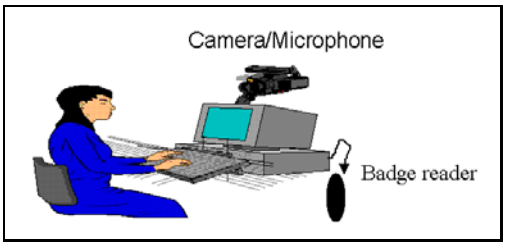



Figure 12: *Multimodal person verification scenario based on face, speech, and lip movements.*

The combination of several modalities for personal identity verification is motivated by the fact that monomodal systems often don't meet the high performance requirements imposed by typical applications, whereas the combination of modalities can improve their performance. Within the

European project M2VTS we have made several contributions in the area of multimodal person verification. These activities are now being continued on the follow-up project BANCA (Biometric Access Control for Networked and e-Commerce Applications) where biometric person authentication is extended for applications over the internet. A typical setting for multimodal access control is depicted in Fig. 12.

**Verification Modalities**

**Acoustic Speaker Verification**. The group has developed and evaluated different speaker verification methods, text-independent techniques based on second order statistical moments and text-dependent method based on hidden Markov models. To increase the robustness of these systems to noisy environments and channel variability we have investigated several methods including cepstral mean subtraction and signal mean subtraction. Furthermore, a technique has been developed to prune the speech data of the accessing person by selecting utterances that are most discriminant for verification. The methods were extensively evaluated on different databases to assess the performance and limitations of the resulting system.

**Visual Speaker Verification**. We have developed a biometric person authentication method based on visual motion information of the face, particularly the lips, while the person is talking. Visual information is extracted by tracking the lips over image sequences and extracting both shape and grey-level information of the mouth region. The verification technique uses a combination of spatio-temporal models based on hidden Markov models (HMM) and appearance based visual models. As discussed in Sec. 3.1, HMMs are extensively used in speech processing and represent a very efficient method to model the statistical variation in both the temporal and the feature domain. We exploit these properties to model the visual appearance of the speaker during speech production.

**Face Verification**. Databases play an important role for the development and evaluation of person verification algorithms like face verification. Despite this fact, there exists no measure that indicates whether a given database is sufficient to test a given algorithm. We have proposed a method that evaluates the complexity of a given database to evaluate if a database is appropriate for the simulation of a given application. Results on four commonly used databases reported in the literature suggest that two of them are insufficient for realistic evaluations of face recognition algorithms.

**Sensor Fusion**

Sensor fusion is a powerful solution to pattern recognition problems involving complementary classifiers and noisy input since it allows the simultaneous use of different information sources. Typical problems which arise in person verification research are the combination of different information sources (frontal face, face profile, facial motion, speech) and different information representations (still, dynamic). The fusion of classifiers is particularly difficult in the case of classifiers exhibiting different performance levels.

**Combination of Classifiers**. The group has investigated and evaluated various methods of sensor fusion in collaboration with the Machine Learning Group, including Support Vector Machines (SVM), multi-linear classifiers, MLP, and C4.5. Experimental results show that the fusion of classifiers increases the performance and outperforms each single verification modality.

**Combination of Acoustic and Visual Speaker Verification** Based on the acoustic and the visual speaker verification techniques we have proposed a multimodal system that combines both modalities, the visual and the acoustic stream, in a single framework. Both information streams are first segmented into words by the application of an acoustic speech recognition

system. Verification techniques based on HMMs are then applied to both streams using word models. The likelihood measure for the two streams are normalised and mapped to a common confidence interval that enable the combination of the two modalities. Results indicate that the performance of the visual sub-system is considerably lower than for the acoustic system. However, the combined system improves the performance of the acoustic system.

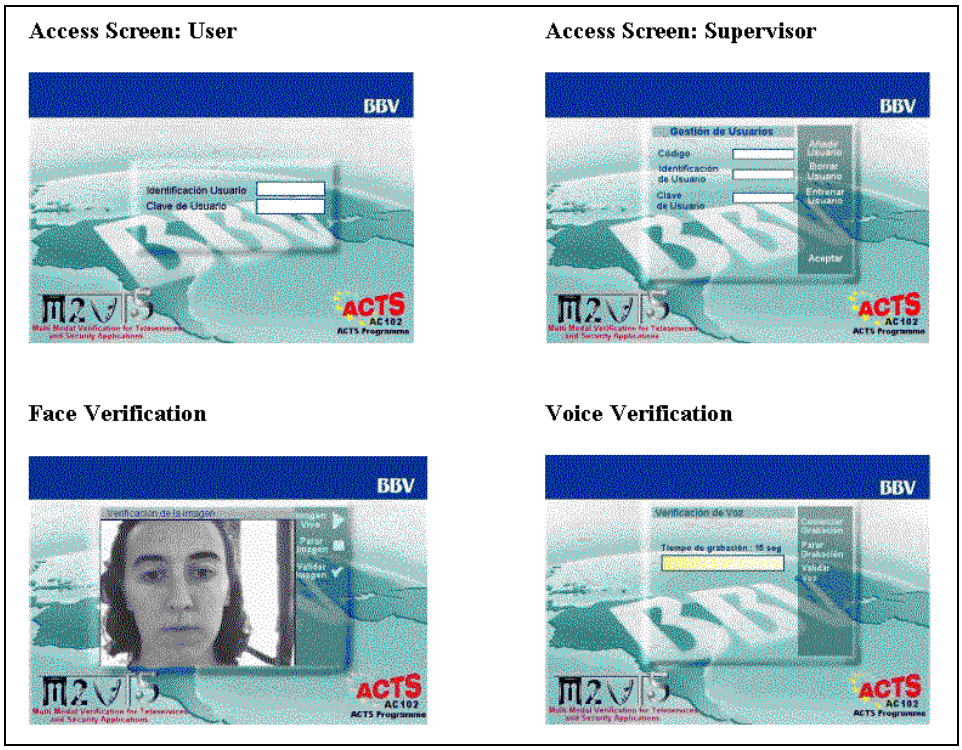



Figure 13: *Multimodal user verification for access control to restricted documents or services on the WWW (tele-banking, tele-shopping, tele-working, etc.). The application interface was developed by Ibermatica SA for Banco Bilbao Vizcaya, integrating our verification technology.*

### Demonstrator and Field Tests

The performance of verification systems tested on laboratory databases often drops considerably when applied to real-world situations. In collaboration with Siemens AG (formerly Cerberus AG) and the University of Neuchatel, we have implemented a verification system prototype including user interface, client registration, training, and verification possibilities. To allow the evaluation of verification performance in realistic applications, the system has been used to collect a field test database, where 22 members of IDIAP were recorded during a period of several months. This database represents a very difficult application scenario which is characterised by a low-cost microphone, many non-native speakers, reverberations, and varying distance and direction of the microphone. Verification experiments have been performed on this database to test the robustness of algorithms under these conditions. Potential applications targeted by Siemens are access control to secure buildings and remote verification of triggered alarms.

Ibermatica SA, has integrated our verification technology and has developed several application prototypes including multimodal verification for cash dispenser access, computer access, and internet access (Fig. 13).

### 3.2.2 Acoustic-Visual Speech Recognition

The performance of speech recognition systems drops significantly in the presence of noise and basically all typical applications are subject to some kind of noise. The objective of this work is to exploit the complementary information present in the visual speech signal (lip-reading) to enhance the performance of speech recognition systems in noisy conditions, as it is naturally done by humans.

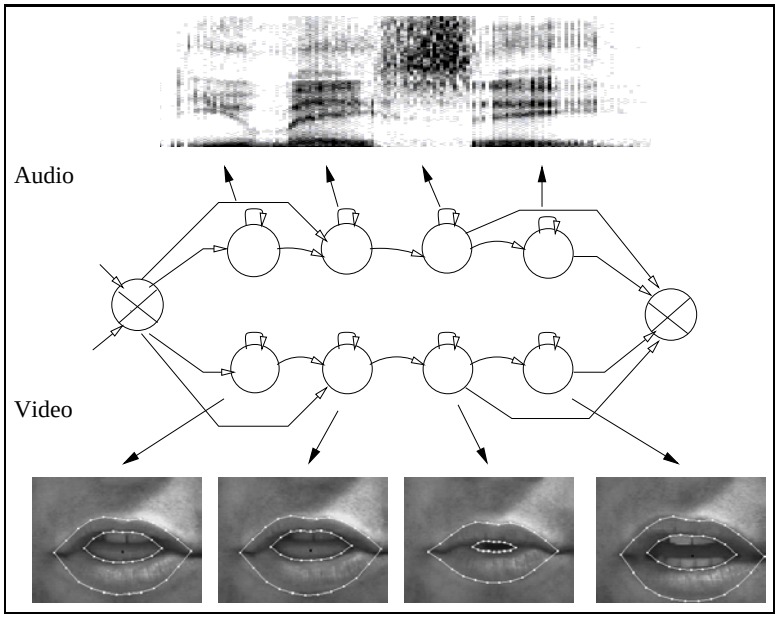



Figure 14: *Acoustic-visual speech recognition. Whereas acoustic features are based on traditional methods (e.g. MFCC) or noise-robust methods (e.g. RASTA, RASTA-PLP), visual features are extracted from the lip contours and the intensities around the mouth. The two information streams are combined using multi-stream hidden Markov models.*

#### Lip-Tracking and Feature Extraction

This work includes research in motion analysis and in the analysis and representation of visual contours. We have developed a technique for the modelling and tracking of deformable objects which was applied to the problem of lip-tracking. The work was concerned with appearance based modelling to enable both the difficult task of tracking and the parameterised representation of deformable objects.

A visual speech recognition system purely based on visual information has been investigated. Visual features are extracted from the results of the lip tracker and represent both contour information of the lips and grey-level intensity information of the mouth based on the appearance based representation. The system has achieved performance levels similar to human lip-readers that have no lip-reading knowledge.

#### Sensor Fusion

Sensor fusion for audio-visual speech recognition is concerned with the fusion of asynchronous, possibly noisy, data. In collaboration with Faculté Polytechnique de Mons (Belgium), we have investigated a method based on the Multi-Stream approach (see Speech Processing Group) which enables the synchronous decoding of audio-visual speech but which can still account for asynchrony between the two modalities. The system has been evaluated for a continuous speech recognition task and has shown to considerably improve the performance of acoustic-only systems when background noise is present, as illustrated in Fig. 15.

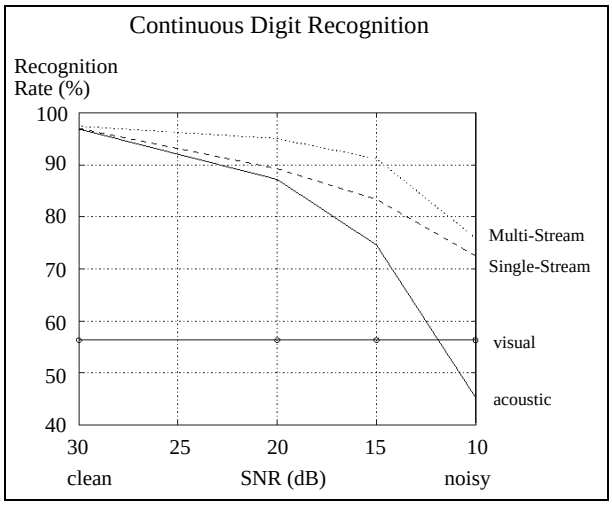



Figure 15: *Recognition rate for acoustic, visual, and audio-visual speech recognition. Audio-visual performance is displayed for traditional fusion at the feature level (Single-Stream) and for fusion based on the Multi-Stream method.*

### 3.2.3 Facial Expression Recognition

The objective of this project is to develop robust and accurate computer vision techniques for the visual analysis and recognition of facial expressions from image sequences. This project benefits from collaboration with the Psychology Department at the University of Geneva, which has considerable expertise in the area of facial action coding and human emotion analysis. We are also collaborating with the Department of Otolaryngology at the Geneva University Hospital that is interested in the automatic evaluation of facial nerves.

Facial expression recognition should not be confused with human emotion analysis as it is often done in the computer vision community. Whereas facial expression recognition is concerned with the classification of facial motion into abstract classes, purely based on visual information, human emotion is the result of many different factors and it's state might (or might not) be revealed through a number of channels of which the face is just one of them (16).

The application of currently available automatic facial expression recognition systems to the analysis of natural scenes is often very restricted due to the limited robustness of these systems and the hard constraints posed on the subjects and on the recording conditions. Our approach aims to overcome these shortcomings by investigating advanced computer vision techniques for the modeling of faces and scenes.

#### FACS Unit Database

We have created a digital facial expression database based on image sequences provided by the Psychology Department at the University of Geneva. This database contains image sequences according to FACS (Facial Action Coding System), a de-facto standard for categorising facial actions independent of emotion. This database has been categorised by a FACS expert and therefore provides a realistic basis for the evaluation and comparison of facial expression recognition techniques with each other and with the human expert.

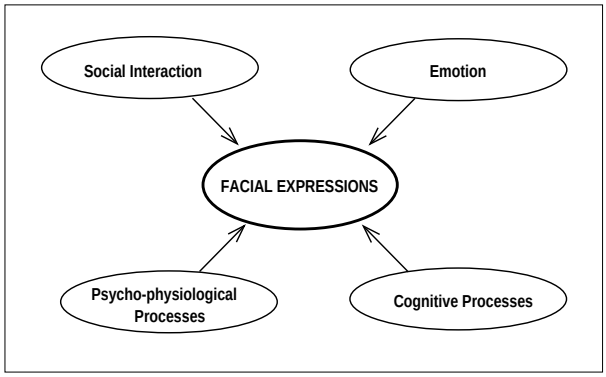



Figure 16: *Mental and physical activities that can elicit facial expressions*

**Recognition of FACS Units and Intensities**

This work concerns the developed and implemented facial expression analysis techniques that are evaluated on the collected FACS Unit Database. We have developed a system that is able to classify facial expression categories according to the FACS standard. A performance comparable to a human expert could be obtained for the single-subject database recorded under controlled conditions. The system is furthermore able to recognise asymmetric FACS units (i.e. units that are only displayed on one side of the face) as well as the intensities of the respective units. Fig. 17 shows two original example images, the same images with the artificial markers removed, and the corresponding facial motion.

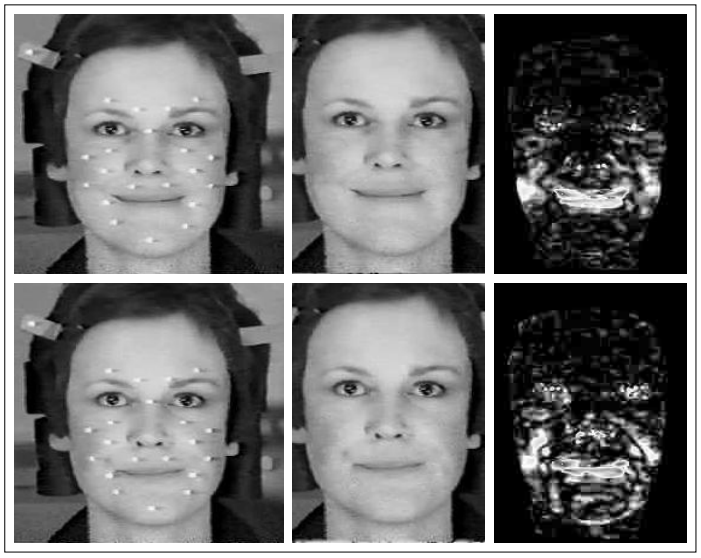


Figure 17: *The left-hand side shows the original marked face, the middle image displays the morphed face without markers, and the right-hand image shows the corresponding facial motion analysis.*

### 3.2.4 X-Ray Image Sequence Analysis

X-ray films showing the side-view of the vocal tract still provide the best dynamic view of the whole vocal tract and provide detailed temporal information about the individual articulators. Many important research results in speech science have been based on such data. In those studies, quantitative information of the articulators was extracted by hand, which restricted the analysis in both the number of samples and the detail of measurement. The automatic analysis of articulatory information would therefore be very beneficial to the scientific community. This project deals with the extraction of articulatory information of X-ray images and is performed in collaboration with the speech processing group.

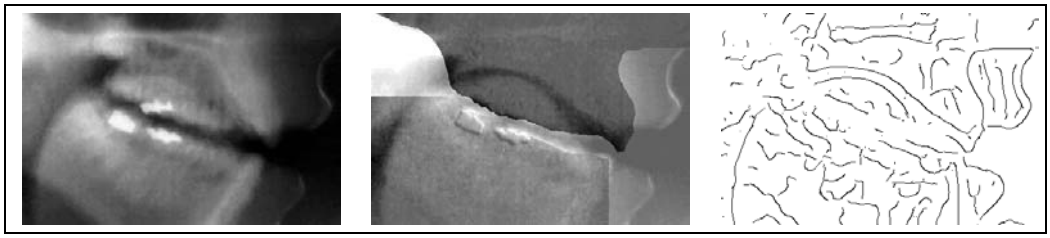


Figure 18: *The left image shows the normalised X-ray image of the side view of the vocal tract. The middle image shows the same image with the upper and lower jaw subtracted. The otherwise hardly visible but important tongue contour can now easily be distinguished. The right image displays the corresponding edge image.*

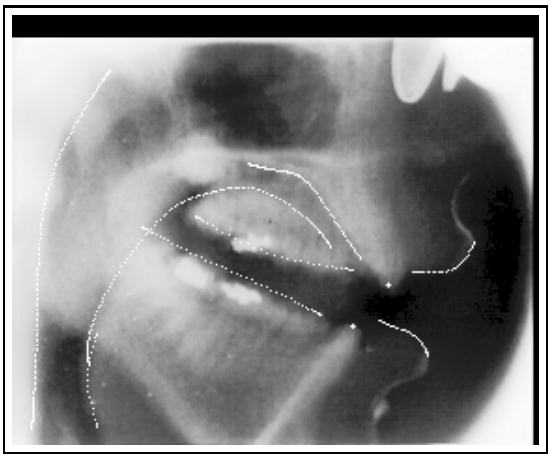


Figure 19: *Results of articulatory parameter extraction from and X-ray image.*

A technique has been developed within the group to track tongue, lips, teeth, and throat in the X-ray film database provided by ATR (Japan), which is probably the largest database of its kind. Our tracking method uses specialised histogram normalisation and a tracking method that is robust against occlusion, noise, and spontaneous, non-linear deformations of the articulators. The tracking robustness is improved by the introduction of temporal constraints, which, however, can cause tracking errors in the case of fast movements. To compensate for this effect, a method that tracks the articulators both forward and backward in time and that optimally combines the results, has been developed. Figure 18 shows some of the processing steps of an X-ray image and Fig. 19 the results of the extracted articulatory features.

### 3.2.5 Document Analysis and Recognition

Document analysis and recognition is concerned with the recognition of machine-printed, hand-printed, and cursively-written documents. Research activities in our group started in 1992 and since then mainly concentrated on hand-printed and cursively-written text recognition.

**Handwritten Character Recognition**

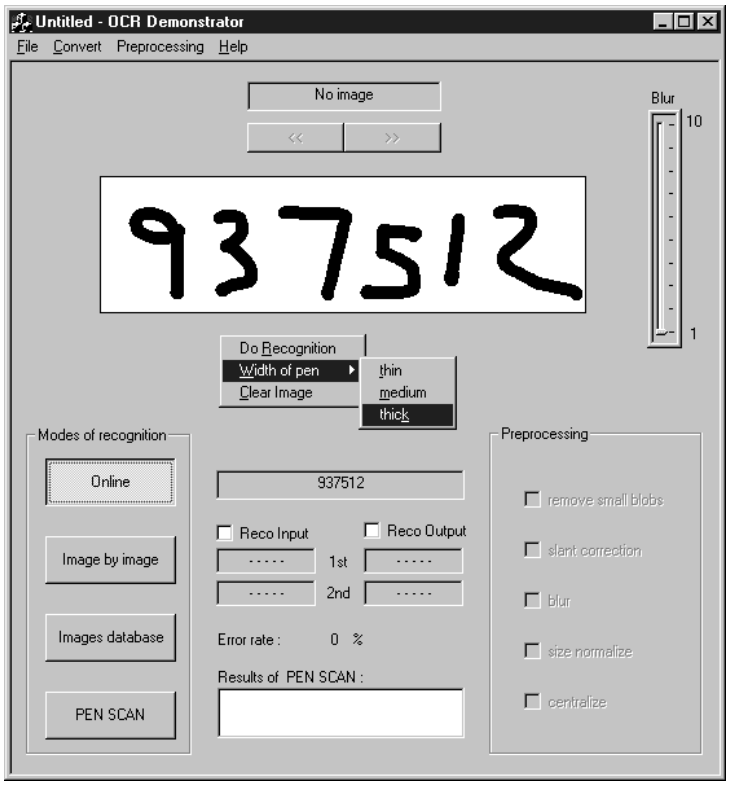



Figure 20: *Handwritten character recognition demonstrator. The system can be used in 4 different modes: recognition of a whole database, recognition of single character of a database (left image), recognition of text scanned by a pen-scanner, or recognition of text written by the mouse (mode shown in the image).*

**Feature Extraction**. Recent work has concentrated on visual feature representations and pre-processing techniques. We have investigated different image transforms for the normalisation of characters and for the alignment between test samples and training samples/models. Other work is concerned with the extraction of relevant features, in particular, methods that are invariant to image transforms corresponding to noise, but which do not reduce class discrimination.

**Classification**. Work in the area of classification has been investigating the use of Support Vector Machines (SVM) for the classification of hand-printed characters. SVM have recently gained much interest and have shown several advantages over alternative classifiers. Experiments were mainly carried out on the NIST Special Databases.

**Pen-OCR System** In the framework of a diploma thesis and in collaboration with the Ecole d'Ingenieurs du Valais we have implemented a printed character recognition system to be used together with a smart-pen. The smart-pen contains a built-in scanner that allows to scan hand-written text that is then transfered and recognised by the character recognition system (Fig. 20).

**Cursive Handwriting Recognition**

Handwriting recognition can be classified into off-line and on-line recognition. Whereas off-line recognition is only based on the image of the written text, on-line recognition is based on pen-input, measured by a device such as a digitising tablet which also provides temporal information. Our activities are in off-line recognition, particularly in cursive handwriting recognition, also called script recognition. Cursive handwriting recognition by machine is very difficult mainly due to the large variability of human writing and due to the fact that the letters are usually connected.

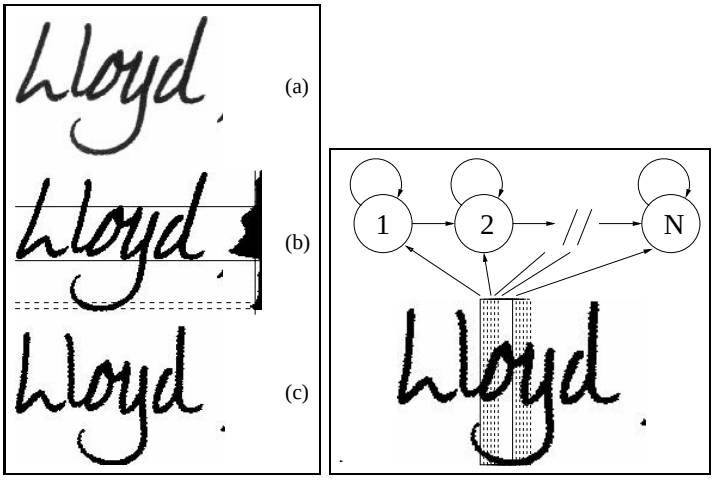



Figure 21: *The core region of the text is found using the horizontal density and the Otsu-method for finding the threshold (left image (b)). The image is further de-sloped and de-slanted (left image (c)). Words are modelled with hidden Markov models using features extracted from a sliding window (right image).*

**Preprocessing**. We are investigating methods for the normalisation of different handwriting styles, e.g. slant correction, slope correction, and core region extraction. In particular, we are interested in methods that are statistically based and that use a minimum of heuristics.

**Feature Extraction**. A large number of feature extraction methods have been proposed in the literature, ranging from simple low level feature extraction methods (e.g. pixel values) to sophisticated methods (e.g. contour models, contour orientation, sub-character units, chain graphs). We are investigating data-driven feature extraction methods that are often more robust than sophisticated techniques. We are also working on methods that allow the seamless combination of the extracted features with hidden Markov models, that are used for handwriting modelling.

**Handwriting Modelling**. One of the fundamental problems in cursive handwriting recognition is the segmentation of characters. This problem is similar to the segmentation of phones in automatic speech recognition. We are investigating methods that address this problem by joint segmentation and recognition using stochastic methods, i.e. hidden Markov models (HMM). The use of HMM also allows the incorporation of multiple knowledge sources such as lexica, language models, and document topic.

### 3.2.6 Image and Video Indexing and Retrieval

Most of the information available today is stored on text, still images, and videos. While today, some of this information is already available in digital form, it is predicted that most of our information sources in the future will be stored on digital media. The Internet and the World Wide Web have mainly been responsible for the world-wide access to this information highway. A general problem in the management of such vast information sources is the task of searching for a particular item that matches a given description. Although text-based search engines are very powerful for searching text documents, they can not directly be applied to the increasingly important audio-visual material. Analysing multimedia documents and finding semantic representations of their content is therefore necessary.

The aim of this project is the content based analysis of images and videos to be used for indexing and retrieval. Content information can be obtained by analysis of the audio track, by the detection and recognition of text present superimposed or present in the scene, or by visual scene analysis methods (e.g. object detection or analysis based on colour, texture or motion). In the framework of the recently accepted European project ASSAVID, we are addressing the particular application of sports video indexing and retrieval.

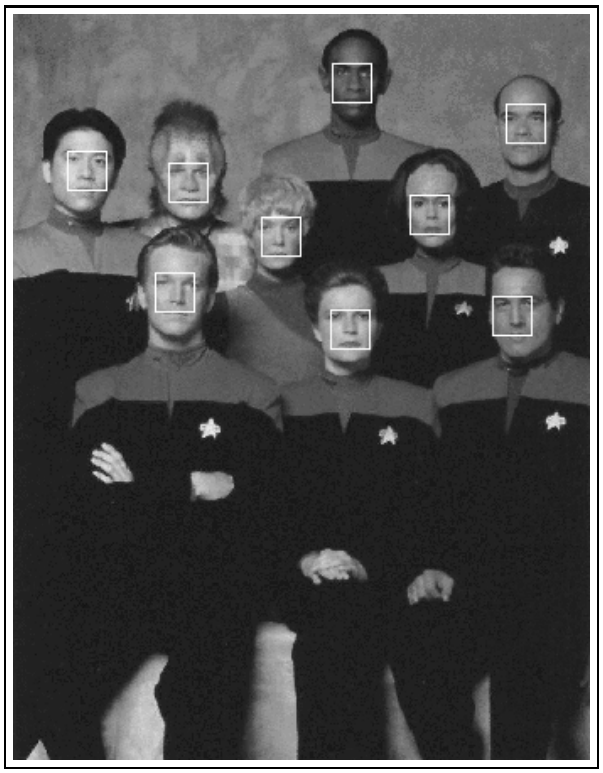


Figure 22: *Example image with face detection results.*

#### Object Detection

Object recognition represents one of the fundamental problems in computer vision. A sub-problem is concerned with the detection of objects of known classes. We have investigated new object detection techniques and have applied them to the problem of face detection, which is a pertinent, yet difficult, task that is required in several vision applications, e.g. face recognition, visual speech recognition,

image and video indexing, video conferencing, and video coding. The task of face detection consists in the analysis of an entire image aiming to detect all faces that appear in the image. A typical example is illustrated in Fig. 22.

In the case of image or video retrieval, object detection algorithms can be used to index documents with respect to their contents of certain objects or they might be used during retrieval to select a certain object and search for this object in the database. Most image and video material contains people and the capability to search for a certain person, by the face for example, would be very beneficial in image or video retrieval.

**Text Detection and Recognition in Images and Videos**

As an important representation form of human being's language, visual texts are widely used in our daily life. The problem of extracting text information from visual clues have attracted wide attention for many years. Great progress has been made in processing printed characters against clean background, such as scanning document pages. The pixels of the text in the scanning result of the document can be easily separated from the background. While, today, more and more information is transformed into digital form, visual text is embedded in many forms of digital media, such as images and videos.


Figure 23: *Example image taken from a sports video frame containing both text superimposed on the video and text that is part of the scene.*

Comparing with the texts in documents, the text in media is in small quantity, but often carries crucial information of the media contents. They usually present important names, locations, brands of the products, scores of the match, date and time, which are helpful information to understand and index these images and videos. However, the text in the images and videos can be superimposed on the arbitrary backgrounds or embedded on the surfaces of the objects in the scene with vary font, size, color, alignment, movement and lighting condition, which makes the text extraction extremely difficult. The aim of research on text detection and recognition in images and videos focuses on finding the proper way to extract different types of text from arbitrary complex images or videos. It will not only extend the application of OCR system into wider multimedia areas but also help people further understand the mechanism of visual text detection and recognition.

### 3.2.7 Research Grants

♢ ARTIST – Articulatory Representations To Improve Speech Technologies

**Funding:** Swiss National Science Foundation

**Duration:** April 1997 – March 1999

**Person involved:** Sacha Krstulović and Georg Thimm

**Description:** This research project aims at using articulatory features in speech recognition (SR) and speaker verification (SV) applications. Such features are believed to lead to significant improvements of SV/SR systems, in accordance with the statements of Liberman's "Motor Theory of Speech Perception", with European ACCOR project's results, and with several other studies.

Subtasks involved in this research include :

- Automatic segmentation of an X-ray video database displaying the vocal tract by means of computer vision techniques. This will provide a set of matched acoustic/articulatory data suitable for the training or validation of acoustic-to-articulatory conversion schemes.
- Implementation of robust acoustic-to-articulatory conversion methods. This will enable the extraction of articulatory features from a sound input, thus making the use of articulatory features compatible with existing SV/SR systems (See Sec. 3.1).
- Use of extracted articulatory feature in SR/SV systems. This will validate the original concept of "Motor Speech Perception" and improve the existing SV/SR applications' performances (See Sec. 3.1).

**Achievements:** We have developed a technique that tracks tongue, lips, teeth, and throat in X-ray image sequences showing the side few of the vocal tract. At the research level, this has required the development of several techniques including specialised histogram normalisation, robust tracking against occlusion and noise, forward-backward tracking, and shape representation and analysis.

♢ FaceX – Facial Expression Recognition through Temporal and Appearance Based Models

**Funding:** Swiss National Science Foundation

**Duration:** October 1998 – September 2000

**Partners:** University of Geneva

**Person involved:** Souheil Ben-Yacoub and Beat Fasel

**Description:** The objective of this project is to develop robust and accurate computer vision techniques for the visual analysis and recognition of facial expressions from image sequences.

The results of this work are important in numerous domains: research and assessment of human emotion (psychiatry, neurology, experimental psychology), consumer-friendly human-computer interfaces, interactive video, and indexing and retrieval of image and video databases. The output of the project will also provide important but missing tools in related research areas such as face recognition, audio-visual speech recognition, lip synchronization, synthesis of talking faces, and model-based image coding (new MPEG standard).

**Achievements:** First, an appropriate facial expression database was created based on image sequences provided by Kaiser and Wehrle from the University of Geneva (FAPSE). Originally, faces were covered with markers which had to be removed using morphological operators. Furthermore, an automatic facial expression analysis system has been implemented that uses difference images (generated by subtracting a neutral face from a given test face) for

extraction and principal component analysis (either PCA or ICA) for representation. We were able to successfully classify facial expressions according to the FACS (Facial Action Coding System) scheme. A performance comparable to a human expert could be obtained for a single-subject database recorded under controlled conditions.

♢ SCRIPT – Cursive Handwriting Recognition

**Funding:** Swiss National Science Foundation

**Duration:** October 1999 - September 2001

**Person involved:** Alessandro Vinciarelli

**Description:** As of today, the capabilities of current cursive script recognition (CSR) systems is far behind the performance of human readers which prevents their use for practical applications that demand high robustness and low error rates.

The main objective of this work is to investigate, develop, and test new CSR methods that could improve the state-of-the-art of current systems.

Our approach is based on the combination of mathematical methods from a number of different research domains including computer vision, statistical pattern recognition, speech recognition, and language modelling. We believe that such a multi-disciplinary approach holds considerable promise to advance research in the field of cursive script recognition. Our approach to address open research issues is as follows:

- Investigation of shape modelling technique used in computer vision and computer graphics for character representation and feature extraction.
- Determination word/character units that are most suitable as basic model units (sub-letters, letters, words, etc) for CSR.
- Application of training and recognition methods based on Hidden Markov Models (HMM).
- Integration of higher level knowledge sources (lexicon, n-gram language model, document topic) in the recognition process.

**Achievements:** During the first few months of the project, a segmentation free script recognition technique based on a sliding window and hidden Markov modelling has been developed. On each frame isolated by the window, a feature extraction process is applied and a sequence of observations is so produced. For each word in the lexicon, a model is trained by concatenating single letter models. The system has been tested over a database composed by words written by a single person. Although the developed system is still in its initial phase, the performance (more than 80%) is already comparable to other results reported in the literature on the same data.

♢ VOCR - Text Recognition for Video Retrieval

**Funding:** Swiss National Science Foundation

**Duration:** December 1999 - November 2001

**Person involved:** Datong Chen

**Description:** Text embedded in images and videos represents an important source of information that can be used for indexing and retrieval. Image and video data often contains text that might appear either superimposed or as part of the scene. Typical examples include news, advertisements, sport, finance, stock market, and geographical maps.

The objective of this project is the investigation and development of algorithms for the detection, segmentation, and recognition of text in images and videos to be used for indexing

and retrieval. Different image properties will be investigated including colour, texture, geometry, and character shape. In addition, the analysis of videos will exploit temporal characteristics of both the scene and the text. An important topic of the project will deal with the combination of evidence acquired my the different modules to perform detection and segmentation. Whereas previous research has treated detection, segmentation, and recognition as three separate problems, that often lead to individual errors, this work will investigate integration methods for all three processes to draw a joint decision driven by the result of the text recognition module.

♢ ASSAVID – Automatic and Segmentation and Semantic Annotation of Sports Videos

**Funding:** European project, 5th Framework Programme, Information Societies Technology, supported by OFES

**Duration:** February 2000 – July 2002

**Partners:** Sony (UK), ACS (I), BBC (UK), University of Firenze (I), University of Surrey (UK)

**Description:** The aim of the project is to develop techniques for automatic segmentation and semantic annotation of sports videos. The problem is relevant to the broadcast industry where the current content annotation relies primarily on the auxiliary information gathered about a broadcast programme during its production. This methodology is appropriate in situations where the material is conceived, scripted, cast and produced under the control of the production department and the auxiliary information can be used for annotation purposes. In less structured environments, particularly in live broadcasts (international competitions, tournaments, Olympic games, etc.) where the content creation is often reactive, or where a transmission feed is taken from another broadcaster, production planning and auxiliary information may be unavailable, or in a foreign language. In this situation, video material may reach the archives with completely inadequate labelling (Olympic games day1) and has to be sorted and annotated laboriously. The project will address this problem and will aim to develop the necessary technology for such multimedia content packaging and devise facilities which will allow a rapid access and search of such material. The level of annotation should be sufficient to enable text-based queries. The target will be to segment the material into shots, and to group and classify the shots into semantic categories (type of sport). To do this, the system will extract information from each shot, based on speech and text recognition, and identify the highlights from the audio track and from visual audience reactions. A training system will then be developed that will then associate these features with a small thesaurus relevant to the subject matter.

♢ BANCA – Biometric Access Control for Networked and e-Commerce Applications

**Funding:** European project, 5th Framework Programme, Information Societies Technology, supported by OFES

**Duration:** February 2000 – July 2002

**Partners:** Matra Nortel Communication (F), Banco Bilbao Vizcaya (E), EPFL (CH), Ibermatica S. A. (E), OSCARD S. A. (F), Thomson-CSF Communications (F), Universite Catholique de Louvain (B), University of Surrey (UK)

**Description:** The objectives of the project is to develop an implement a complete secured system with enhanced identification, authentication and access control schemes for applications over the Internet such as tele-working and Web-banking services. One of the major innovations of this project will be to obtain an enhanced security system by combining classical security protocols with robust multimodal verification schemes based on speech and image. The project includes the following objectives:

- development of scalable and robust multimodal verification algorithms
- development of scalable classifier combination techniques
- design and implementation of an overall secure architecture including security protocols adapted to biometrics
- development of three demonstrators: tele-working, home-banking, and ATM.

## 3.3 Machine Learning Group

**Group Leaders: Samy Bengio and Eddy Mayoraz**

In its broad sense, machine learning means solving problems from samples, usually through the inference of a computational model. The technologies studied, elaborated and experimented in our group are of various natures (based on statistics, on logical concepts or on geometrical considerations) and most of them require heavy optimization techniques. They range from *Bayesian learning* to *support vector machines*, including *neural networks*, *decision trees*, *logical analysis of data*, and *hidden Markov models*.

The activity of our group is always application driven. The learning techniques involved are not studied for their own sake, but are always targeted towards a specific application. As illustrated in Figure 3, the research themes in the Machine Learning Group are articulated along two lines :

- providing technological support for the other two groups;
- investigating new fields for applications of the technology at hand.

The typical characteristics of learning tasks for classification problems arising in perceptual AI are:

- a large number of classes: 10 digits, 26 letters, 30 or 60 phonemes, $N$ potential users of a system with access security, etc.
- a large number of attributes resulting from a preprocessing step (in speech the LPCC, the MFCC or the RASTA coefficients are around 40 to 60; in vision, it is usual to work with hundreds of coefficients),
- a huge number of data patterns (several thousands, or tens of thousands),
- a highly noisy or non-stationary environment.

Thus, our research effort as support for the Speech Processing Group and the Computer Vision Group involves being aware of the latest learning techniques (e.g. *mixtures of experts*, *support vector machines*, *transduction*) and working on strategies to adapt any efficient learning technique to very large problems. The decomposition of large problems into a series of simpler subproblems provides a general approach that does not depend on the learning technique to be used and which has been shown to be quite efficient in practice. *Logical analysis of data* (LAD) is a recent method aiming at extracting knowledge from data in a form that is as easy to understand as possible. This theory has emerged in the operations research community and is still not well known in the learning community. Almost all the work related to LAD is so far carried out at RUTCOR (Rutgers Center for Operations Research) and at IDIAP, involving a fruitful collaboration between the two research groups.

On the prospecting side, we have lately initiated two new areas of research, namely in time series prediction and spatial data analysis. In both cases, the motivation is to demonstrate that some recent modeling or learning tools can provide an improvement in comparison with the methods usually used in these fields.

### 3.3.1 Divide and learn

A good way to handle large learning tasks is to decompose them into a series of smaller and/or simpler learning tasks called *subtasks* or *subproblems*. This implies a strategy for the *decomposition* defining each subtask and a *reconstruction* technique specifying how the answers of each classifier trained on the subtasks are recombined to provide the global answer. The advantages of such approaches are numerous:

- each subproblem is simpler than the global problem;

- some redundancy in the decomposition makes the global model more robust;
- both in the training stage and in its usage, the process can easily be parallelized.

### $K$-class classification turned into 2-class classification

Some learning techniques are designed essentially for the resolution of dichotomies, i.e. 2-class classification problems. Others are more general but scale up badly with the number of classes. A fruitful approach to the resolution of problems with a large number of classes, consists in decomposing the general problem into many subproblems involving two classes only. This approach adds two advantages to the ones listed above:

- a larger variety of learning methods can be used to solve the subproblems;
- since a class in a subproblem envelopes usually several classes of the whole problem, the sample for each class (in a subproblem) is much larger.

Figure 24 illustrates this idea on a simple example of 10 classes decomposed into 4 dichotomies. Each dichotomy is represented by a closed line whose inside and outside define the two classes. This example

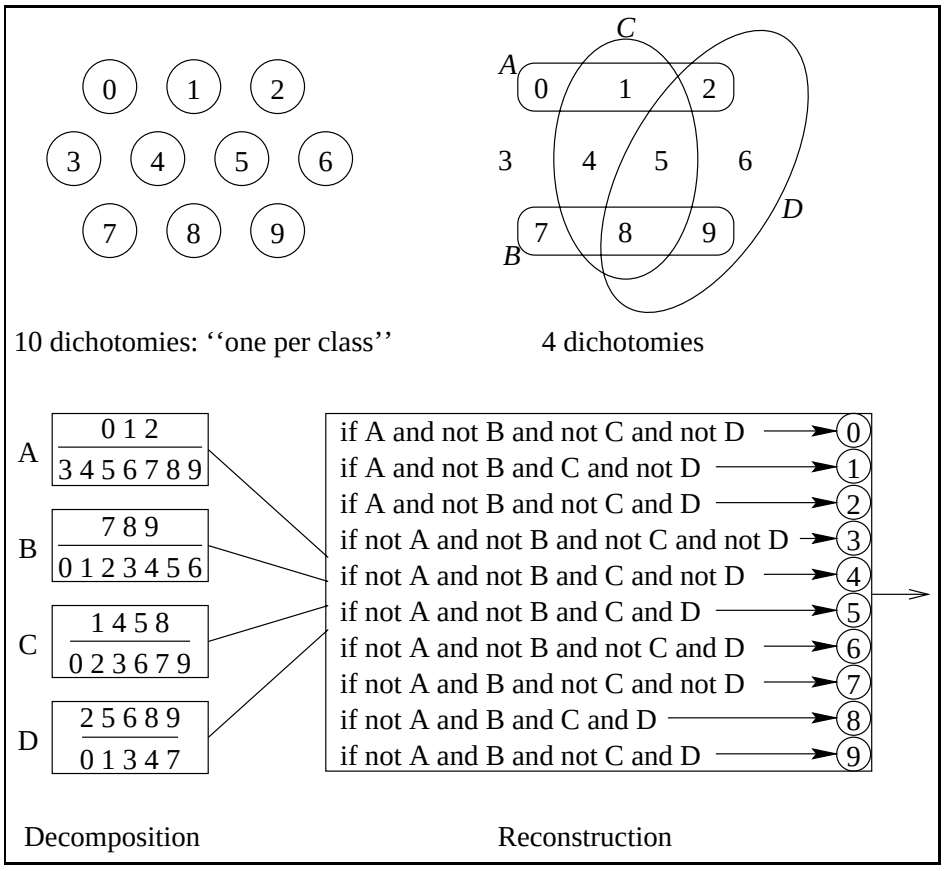



Figure 24: *Decomposition of a 10-class problem into 10 dichotomies (up left) and 4 dichotomies (up right).*

can be refined in many respects. The rule-based reconstruction pictured in the lower part of figure 24 can be advantageously replaced by an algebraic expression (vector-matrix product). An alternative would be to consider the reconstruction as a new classification problem for which any method could be

investigated (stacking). The number of subtasks can be enlarged on purpose in order to get additional robustness from the reconstruction. Moreover, in the example of figure 24, each class is involved in each subtask to determine either positive or negative examples. But generally, some classes can be ignored by some subtasks (e.g. a subtask is defined for each pair of classes, discriminating between these two and ignoring the others).

Recently, our group has carried out important work in this field and is now integrating the obtained results into related research topics. Section 3.3.2 shows a concrete example of the applicability of these principles.

**Mixture of experts**

In a more general framework, the decomposition/reconstruction principle is used to split a complex problem into simpler ones without being restricted to the reduction of the number of classes in the subproblems. Moreover, instead of determining the reconstruction strategy only on the basis of the outputs of the learners resolving the subtasks, it can depend also on the original inputs of the subtasks. This means that the reconstruction varies with the location of the global inputs in the input space.

One model known as *mixture of experts* and proposed in 1991 by Jacobs, Jordan, Nowlan and Hinton, is exactly of that form. Often presented as a generalization of neural networks, it is composed of *experts* realizing the subtasks and a *gating* determining the coefficients of a linear recombination of the outputs of the experts, as illustrated in Figure 25.

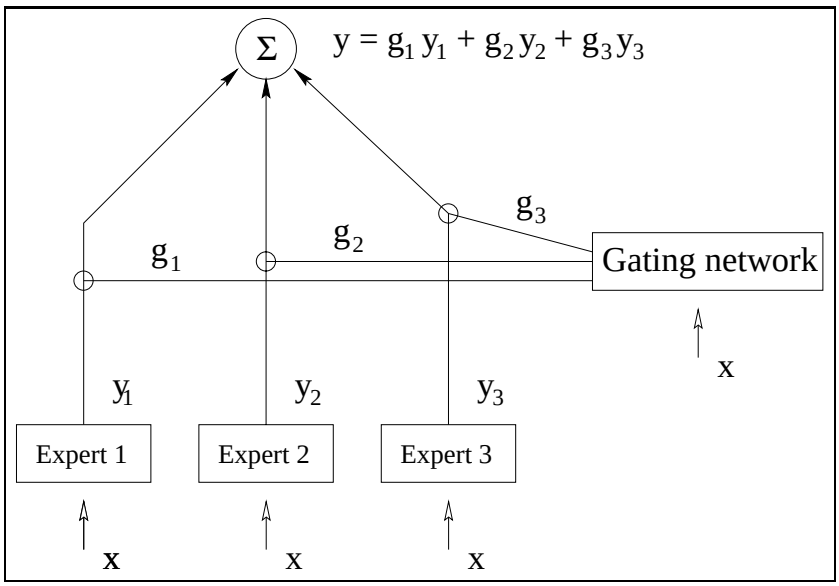



Figure 25: *A mixture of three experts.*

**Expert fusion**

When a classification problem is decomposed into several subproblems (see above), it has been demonstrated that it is often interesting to use different learning techniques for the resolution of the subproblems. This also holds for simple classification problems. Even in the case of a 2-class classification problem, several subproblems can be created. This can be based on different features (see 3.2.1), on a resampling of the training data (bagging, boosting, arcing), or using different types of models . The problem is then to fuse different decisions coming from each classifier into one final decision. In the most interesting case, the decision of each classifier is not just hard decision of class membership, but

it also includes a confidence measure or a probability for this membership.

We carried out some common work with the Computer Vision group in multi-modal speaker identification systems along this line. Different classifiers based on different data (face, lip motion, voice) representing users of the system had to be fused into a single answer. This collaboration was part of the project M2VTS described in Section 3.2.1.

### 3.3.2 Learn and understand what you learn

In practical applications, it is of course desirable that the model provided by an automated learning system is reliable. But in many applications, in particular those related to human sciences (e.g. medical, social, economical issues), it is even more important for the model to be easily understandable for a human expert of the field. Indeed, a physician using computer assisted diagnosis system will be much more inclined to trust such a system if it can give, for any of its answers, a simple justification that can be expressed in a language that makes sense to the physician.

During the eighties, expert systems were very popular tools in AI. They were based on hierarchies of rules, simple and easy to understand. A vestige of these old days are the decision trees, which present many advantages: they are easy to train, after some adequate pruning procedures they provide good error rates and they result in a sequence of simple tests, each being easy to understand.

However, while the human brain is comfortable with IF-THEN-ELSE rules, it is not common to cascade more than two or three of them. A more appealing approach, as far as human understanding is concerned, consists in enumerating a long list of indicators (or patterns or syndromes) that are not strict rules, but just arguments in favor or against a particular diagnosis. These indicators, are not embedded into a hierarchical structure, on the contrary, they are aggregated in a single weighted sum. Given a new case, if many patterns in favor of diagnosis $A$ are active while only few patterns in favor of other conclusions are active, then the diagnosis $A$ is chosen. Logical Analysis of Data (LAD) is a new learning technique which is based on this idea. The elaboration, development and extension of this method is mostly done at RUTCOR and at IDIAP.

A typical LAD session is composed of several sub-steps. The first of these consists in transforming the input data, which can be presented in arbitrary format (numerical continuous, discrete ordered and unordered) into binary format, the one used in all the subsequent LAD operations. Our group has developed significant activity in the optimization of this procedure, whose computational complexity was previously quadratic in the number of data, and thus intractable for large datasets. A new binarization algorithm has been developed with complexity in the order of $O(n \log(n))$, where $n$ represents the number of data. This result broadens the applicability of LAD.

Until now, LAD is capable of handling problems where the number of possible different diagnoses is limited to two, i.e. where the answer of the system can be either yes or no. These are usually called 2-class, or binary problems. The set of patterns (indicators) produced must in this case be able to sufficiently distinguish among cases of two types. At IDIAP, we are involved in extending LAD to problems with several possible classes. Different approaches are possible:

- Use the decomposition techniques referred in Section 3.3.1, and blindly apply LAD for solving each one of the subproblems;
- Integrate those decomposition principles into the core of LAD. In this way, it is possible to better exploit the internal functioning of the method and to produce a more compact and coherent set of indicators. However, this demands a restructuring of the pattern generation procedure.

### 3.3.3 Time series prediction and modeling

The set of problems addressed in machine learning has a coarse division into two classes: classification and regression, depending whether the output of the system takes its values into a finite unordered set or an ordered set. A first refinement of this taxonomy is obtained with the distinction between problems with static behavior (the output of the underlying system depends only on its input) and

those with dynamic behavior (the output of the system at a given time depends on its current input as well as on the input/output history). In a first approximation, one could argue that the second type can be reduced to the first type by expending the input of the system to a window on the most recent inputs. However, this is not suitable for many applications, as it is too sensitive to time-scale variation, and to insertion and deletion of short events.

A newly started project on Time Series Prediction (TSP) is a prospecting activity of our group. In this project, the usability for general TSP problems of well-mastered technology in speech processing (hidden Markov models (HMM), hybrid HMM models and neural networks) is evaluated. Mixtures of neural network experts have lately been shown to be very efficient for certain types of time series (forecast of electricity demand). The use of these models is further investigated for other types of time series (financial or environmental). Classically, in a mixture of expert model, every experts as well as the gater are neural networks. We are also considering models where the gater is an HMM.

### 3.3.4 Spatial data analysis

A second prospecting activity of our group is devoted to the analysis of spatial data. The goal of this research is to exploit and adapt methodologies from the field of learning algorithms for the analysis of environmental spatial data, where data is usually highly spatially non stationary.

### 3.3.5 Research Grants

♢ Divide and Learn

**Funding:** Swiss National Science Foundation, FN 21-45621.95 and FN 55688.98

**Duration:** April 96 – March 00

**Partners:** Swiss Federal Institute of Technology (EPFL)

**Person involved:** Perry Moerland

**Description:** The aim of this project is the study and the extension of the mixture of experts model. This model adaptively partitions the input space (using a gating network) and attributes local experts to these regions. This model has shown to be a powerful tool for dealing with classification and regression problems. The goal of this project was to better understand the influence of the choice of the gating network and develop new methods for learning the parameters of "mixture of experts"-like models

**Achievements:** Research within this project has first been focused on classification problems and the choice of the gating network using methods for density estimation, especially mixture models. As a first step, we have studied the recently proposed mixtures of latent variable models which are a more flexible (in terms of computational complexity and number of parameters) alternative for Gaussian mixture models (GMMs) on the problem of input density estimation. These mixtures of latent variable models often outperformed GMMs in a series of experiments on large number of data sets and also showed very good results when incorporated in so-called Bayes classifiers. However, including these mixture models as a gating network in a mixture of experts did not lead to better results than the ones with standard mixtures of experts (with a single or multi-layer perceptron gate).

We also explored the link between mixtures of experts and a method called *boosting*. Boosting is a method that iteratively generates experts trained on the data that turned out to be difficult for the previous experts. The resulting model is a weighted vote (with fixed weights) over the outcome of the different experts. Boosting has been shown to lead to highly accurate classifiers. An alternative for using fixed weights when combining the expert outcomes, is to allow the combiner weights to be input-dependent. This results in a final model that is close to a mixture of experts. We developed a principled extension of boosting for learning such a dynamic combination of experts. Experimental results indicate

that our approach can lead to results that are at least as good as with boosting (in terms of classification error) but with a considerably lower number of experts.

As a side result of the work on mixtures of latent variable models, we have also been looking into the problem of non-linear feature extraction. A recent method for extracting interesting features from given data in a non-linear way is so-called *kernel principal component analysis* which first non-linearly maps the data in a very high-dimensional space and then extracts features in a linear way. This can be done without mapping the data explicitly by the so-called kernel trick also used in support vector machines and involves the eigendecomposition of a $n \times n$ matrix for $n$ data points. We developed an efficient on-line learning algorithm that allows the application of kernel PCA on large data sets ($n > 10,000$). The extracted features have been used as the inputs of simple linear models and show very good classification performance.

♢ GLAD – Generalization of LAD

**Funding:** Swiss National Science Foundation, FN 2000-053902.98/1

**Duration:** November 98 – October 2000

**Partners:** Swiss Federal Institute of Technology (EPFL)

**Person involved:** Miguel Moreira

**Description:** This project is about the generalization of Logical Analysis of Data (LAD) into a method capable of handling classification problems with large databases. LAD has been shown to be a very efficient machine learning technique for several types of databases. However, so far it is limited to classification problems with two classes only. Moreover, the algorithms available for the resolution of each step of the method scale up very badly with the size of the database (number of data items and number of attributes). In particular, the method designed to solve the first phase of the process (the binarization phase) is quadratic in the number of data, and thus is not usable for problems with more than a couple of hundreds of data items.

In this project, several solutions to generalize LAD to $K$-class classification problems are proposed. New algorithms to make the whole process suitable for large scale problems are developed. For example, a new algorithm solving the binarization for a problem of $n$ data in $O(n \log(n))$ is designed.

**Achievements:** The main achievement of 1999 concerns the creation and improvement of a new procedure for data binarization. The binarization procedure consists mainly in the creation of a good set of binary attributes based on the original attributes, which allows to map every example presented into its binary image. Two (opposing) measures are used to qualify a binary mapping: 1) The size — the smaller the better, for complexity reasons. Many binary attributes generate large binarized data sets. 2) The consistency — a sufficient number of binary attributes is necessary to avoid excessive information losses. The longer the binary mapping, the easier it is to distinguish among different cases, which is important especially if they belong to different classes. The approach used previously to find the binary mapping worked by construction. After enumerating an extensive list of all possible binary attributes, it iteratively chose from the list the best binary attribute of the moment and added it to the solution, until a suitable binary attribute set was found. The newly proposed approach, on the contrary, considers the extensive binary attribute list as the initial solution and eliminates redundant elements iteratively, until the quality of the set is judged appropriate. The efficiency of this approach lies mainly in the way in which the redundancy of each element is calculated.

The elaboration of a new version of the LAD software and corresponding documentation, integrating all the recent research developments and with improved modularity, is another significant result of 1999.

♢ ZEPHYR – Time Series Prediction with Hybrid Markov Models

**Funding:** Swiss National Science Foundation, FN 21-50744.97

**Duration:** January 98 – December 99

**Partners:** Swiss Federal Institute of Technology (EPFL)

**Person involved:** Frédéric Gobry

**Description:** Hidden Markov Models (HMM) have been used extensively and very successfully for speech processing for the past 20 years. Hybrid models mixing adequately HMMs and artificial neural networks are powerful tools for speech recognition. The aim of this project is to study in what extent HMMs and hybrid models can be used for time series prediction. The mixture of experts (MEs) models will also be considered, as well as hybrid models involving MEs.

This study targets different types of time series, from very chaotic ones (financial series) to more structured ones (economical series such as the prevision of electrical demand), including series where the phenomena to be triggered are sparse (floods or avalanches forecasting). For each of these applications, we are considering the most adequate model, combining HMM or mixture of experts models with some appropriate machine learning methods (ARMA models, neural networks, decision trees, support vector machines).

**Achievements:** Hidden Markov models have been experimented on financial time series and on artificial data, both with linear auto-regressive models and multi-layered perceptron (MLP) for the modelling of the emission probability of each hidden state of the HMM. Two training algorithms were implemented, one based on the Viterbi method and the other based on the Expectation-Maximization method. On a first series of experiments, the results were not significantly better than those obtained with a single MLP, and we concluded that the time series used were not composed of multiple regimes.

Other experiments were carried out on physiological time series (heart beat, volume of breath, oxygen level in the blood of a sleeping subject). Interestingly enough, the hybrid model HMM/MLP easily found a segmentation of the time sequence, but the latter was not directly related to the sleep cycles as determined by a specialist. However, the total error rate of the prediction was extremely good.

♢ CARTANN – Cartography by Artificial Neural Networks

**Funding:** Swiss National Science Foundation, FN 2100-054115.98

**Duration:** January 99 – January 2001

**Partners:** Lausanne University (prof. Michel Maignan)

**Person involved:** Mikhael Kanevski and Nicolas Gilardi

**Description:** This work addresses a series of basic research items of spatial data analysis:

- highly non stationary spatial processes,
- cartography of distribution functions, as opposed to cartography of the mean value,
- user and data-driven parameterization for the discrimination between a stochastic trend and auto-correlated residuals,
- cartography of stochastic deviations related to advection-diffusion models.

Final solutions proposed for the resolution of geostatistical problems will mostly be hybrids involving ANNs and other learning methods (such as support vector machines and kernel ridge regression) to extract the general trends, together with classical approaches of geostatistics such as kriging estimations and simulations to estimate the residuals of the learning algorithm predictions if necessary.

**Achievements:** Some experiments were carried out using support vector machines (SVM) to discriminate between two classes of pollution in the Lake of Geneva (cadmium below or above a threshold). Figure 26 illustrates the effect of the parameter $\sigma$ associated to RBF kernels of SVM. Regression experiments have also been done using support vector regression, kernel ridge regression and Bayesian kriging to predict the the level of the cadmium in the Lake of Geneva.

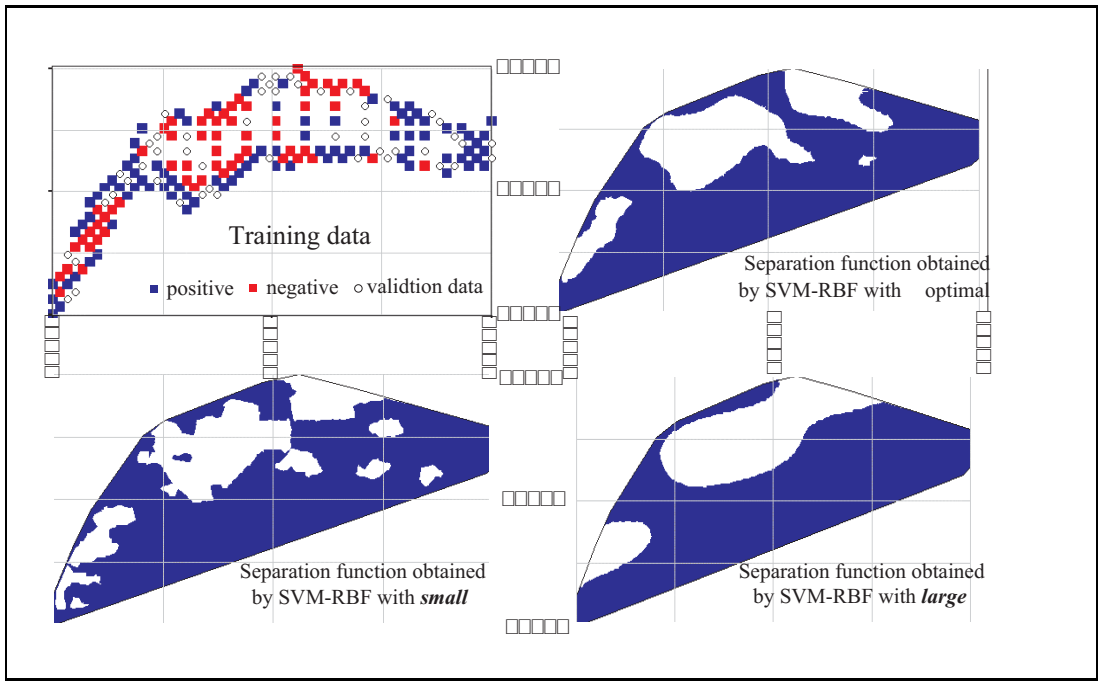



Figure 26: *Cadmium level in the Lake of Geneva.*

The data of the training set are plotted in the upper-left picture. The results obtained with the optimal $\sigma$ are displayed in the upper-right picture, while the lower-left and lower-right pictures illustrate two extreme cases of over-fitting ($\sigma$ too small) and over-smoothing ($\sigma$ too large).

# 4 Educational Activities

## 4.1 Current Ph.D. Theses

- **Ph.D. Candidate:** Giulia Bernardis
  **Supervisor:** Prof. Hervé Bourlard and Dr. Martin Rajman (EPFL)
  **Research topic:** Speech Recognition
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Datong Chen
  **Supervisor:** Dr. Jürgen Lüttin
  **Research topic:** Text Recognition for Video Retrieval
- **Ph.D. Candidate:** Beat Fasel
  **Supervisor:** Dr. Jürgen Lüttin
  **Research topic:** Facial Expression Recognition
- **Ph.D. Candidate:** Nicolas Gilardi
  **Supervisor:** Prof. Michel Maignan (UNIL) and Dr. Samy Bengio
  **Research topic:** Cartography using neural networks
  **University:** Lausanne University
- **Ph.D. Candidate:** Hervé Glotin
  **Supervisors:** Prof. Hervé Bourlard and Dr. F. Berthommier (ICP, Grenoble)
  **Research topic:** Coupling of CASA and Multistream recognition
  **University:** INPG, Grenoble
- **Ph.D. Candidate:** Frédéric Gobry
  **Supervisor:** Dr. Eddy Mayoraz and Prof. Hervé Bourlard
  **Research topic:** Time Series Prediction with Hybrid Markov Models
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Astrid Hagen
  **Supervisor:** Prof. Hervé Bourlard and Dr. Andrew Morris
  **Research topic:** Multistream Speech Recognition
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Christopher Kermorvant
  **Supervisor:** Prof. Hervé Bourlard
  **Research topic:** Robust Speech Recognition
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Sacha Krstulović
  **Supervisor:** Prof. Hervé Bourlard and Prof. Martin Hasler (EPFL)
  **Research topic:** Using Articulatory Features for Speech Recognition / Speaker Verification

**University:** EPFL, Lausanne

- **Ph.D. Candidate:** Perry Moerland
  **Supervisor:** Dr. Samy Bengio and Prof. Wulfram Gerstner (EPFL)
  **Research topic:** Mixtures of experts
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Miguel Moreira
  **Supervisor:** Dr. Samy Bengio and Prof. Alain Hertz (EPFL)
  **Research topic:** GLAD – Generalization of LAD
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Bojan Nedić
  **Supervisor:** Prof. Hervé Bourlard
  **Research topic:** Speaker verification
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Todd Stephenson
  **Supervisor:** Dr. Andrew Morris and Prof. Hervé Bourlard
  **Research topic:** Bayesian Networks applied to speech recognition
  **University:** EPFL, Lausanne
- **Ph.D. Candidate:** Alessandro Vinciarelli
  **Supervisor:** Dr. Jürgen Lüttin
  **Research topic:** Cursive Handwriting Recognition
- **Ph.D. Candidate:** Katrin Weber
  **Supervisor:** Prof. Hervé Bourlard
  **Research topic:** Multichannel Speech Recognition
  **University:** EPFL, Lausanne

## 4.2 Student Projects

- **Trainee:** Pavel Balabko
  **School:** Ecole Polytechnique Fédérale de Lausanne, pre-doctoral school
  **Subject:** Speed/music discrimination based on modulation spectrum
  **Duration:** April 1999-June 1999
  **Supervisors:** Prof. Hervé Bourlard
- **Trainee:** Frédéric Bressoud
  **School:** Ecole d'Ingénieurs du Valais
  **Subject:** Personal Voice Dialling over PC
  **Duration:** October 1999-January 2000
  **Supervisors:** Frank Formaz
- **Trainee:** Antonio Campanile
  **School:** Ecole Polytechnique Fédérale de Lausanne

**Subject:** Mixtures of experts for financial data

**Duration:** April 1999-June 1999

**Supervisors:** Perry Moerland and Prof. Wulfram Gerstner (EPFL)

- **Trainee:** Eric Grand

  **School:** Ecole d'Ingénieur du Valais

  **Subject:** Demonstrator for handwritting character recognition using Support Vector Machines

  **Duration:** May 1999 – June 1999 and October 1999 - January 2000

  **Supervisor:** Dr. Jürgen Lüttin

## 4.3 Lectures

- **Lecturer:** Prof. Hervé. Bourlard

  **Title:** Decision, estimation, and statistical pattern recognition – Application to Speech Recognition

  **Location:** Pre-doctoral school for Computer Science and for Communication Systems, EPFL, Lausanne

  **Duration:** one semester from March 12 to June 18, 1999

- **Lecturer:** Prof. Hervé. Bourlard (together with Dr. Martin Rajman and Jean-Cedric Chapellier, EPFL)

  **Title:** Traitement Automatique du Langage

  **Location:** undergraduate course, department of computer science, EPFL, Lausanne

  **Duration:** summer 1999 semester

- **Lecturer:** Prof. Hervé Bourlard

  **Title:** Speech Processing for Multimodal Interfaces

  **Location:** EPFL Postgraduate School on Multimodal Interfaces

  **Duration:** March 26, 1999

- **Lecturer:** Prof. Hervé Bourlard

  **Title:** Voice-Based Systems

  **Location:** EPFL Postgraduate School on Intelligent Agents

  **Duration:** May 5, 1999

- **Lecturer:** Prof. Hervé Bourlard

  **Title:** Research in the area of multimodal interaction

  **Location:** IIMT MBA, Fribourg University

  **Duration:** September 17, 1999

## 4.4 Examinations

- **School:** Université Pierre et Marie Curie, Paris VI

  **Subject:** Habilitation à diriger des recherches

  **Expert:** Prof. Hervé Bourlard

  **Candidate:** Dr. Marie-José Caraty

  **Title:** Lar reconnaissance vocale et son mentor: l'évaluation

  **Date:** February 1, 1999

# 5 Other Scientific Activities

## 5.1 Editorship

- **Name:** Prof. Hervé Bourlard
  **Function:** Editor-in-Chief
  **Journal:** Speech Communication
- **Name:** Prof. Hervé Bourlard
  **Function:** Action Editor
  **Journal:** Neural Network
- **Name:** Dr. Eddy Mayoraz
  **Function:** Co-Editor of special issues
  **Journal:** Annals of Mathematics and Artificial Intelligence

## 5.2 Scientific Committees Membership

- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Conference:** European Symposium of Artificial Neural Networks (ESANN)
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Conference:** ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Conference:** ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Society:** International Association for Cybernetics
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Conference:** Neural Information Processing Systems (NIPS)
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Scientific Committee
  **Conference:** IEEE Neural Network Signal Processing Society
- **Name:** Prof. Hervé Bourlard
  **Function:** Member of the Administration Committee
  **Conference:** European Association for Signal Processing (EURASIP)
- **Name:** Dr. Eddy Mayoraz

**Function:** Member of the Scientific Committee

**Conference:** European Symposium of Artificial Neural Networks (ESANN)

- **Name:** Dr. Eddy Mayoraz

  **Function:** Member of the Program Committee and organizer of 2 special sessions on artificial neural networks

  **Conference:** 5th International Symposium on Artificial Intelligence and Mathematics

- **Name:** Dr. Georg Thimm

  **Function:** Current Events Editor

  **Journal:** Neurocomputing

## 5.3 Organization of Conference

- **Title:** 1999 IDIAP Symposium

  **Location:** Centre du Parc, Martigny

  **Date:** June 4, 1999

  **Organizer:** IDIAP

## 5.4 Short term visits

- **Location:** Intl. Computer Science Institute (ICSI), Berkeley, CA, USA

  **Visitor:** Hervé Bourlard

  **Date:** March 9-12, 1999.

- **Location:** Department of Medical Informatics at Graz University of Technology, Graz, Austria

  **Visitor:** Mikko Kurimo

  **Date:** June 14-18, 1999.

- **Location:** IRISA Rennes, France

  **Visitor:** Johnny Mariéthoz

  **Date:** from July 16 to August 12, 1999.

- **Location:** Royal Holloway University of London, Egham, United Kingdoms

  **Visitor:** Nicolas Gilardi

  **Date:** from August 24, 1999 to September 24, 1999

- **Location:** Intl. Computer Science Institute (ICSI), Berkeley, CA, USA

  **Visitor:** Hervé Bourlard

  **Date:** August 16-20, 1999.

- **Location:** Symposium: L'art et le virtuel, Sion, Switzerland

  **Visitor:** Beat Fasel

  **Date:** October 12, 1999.

- **Location:** Sheffield University, Sheffield, UK

  **Visitor:** Hervé Bourlard

  **Date:** December 2, 1999.

- **Location:** AT&T, Menlo Park, CA, USA
  **Visitor:** Samy Bengio
  **Date:** December 6 and 7, 1999.
- **Location:** LORIA, Nancy, France
  **Visitor:** Sacha Krstulović
  **Date:** December 6-10, 1999.
- **Location:** Intl. Computer Science Institute (ICSI), Berkeley, CA, USA
  **Visitor:** Hervé Bourlard
  **Date:** December 10 and 11, 1999.
- **Location:** Institut de la Communication Parlée (ICP), Institut National Polytechnique, Grenoble, France
  **Visitor:** Hervé Glotin
  **Date:** 40% of the year.

# 6 Events and Presentations

## 6.1 Scientific Presentations

- **Event:** Neural Information Processing Systems, Denver, CO, USA, November 29 - December 4, 1999

  **Speaker:** Samy Bengio

  **Title:** Modeling High-Dimensional Discrete Data with Multi-Layer Neural Networks

- **Event:** EPFL-Mantra Seminar, Lausanne, Switzerland, December 12, 1999

  **Speaker:** Samy Bengio

  **Title:** Modeling High-Dimensional Discrete Data with Multi-Layer Neural Networks

- **Event:** Second International Conference on Audio-Visual Biometric Person Authentication, Washington D.C., USA, March 22-23, 1999

  **Speaker:** Souheil Ben-Yacoub

  **Title:** Multi-Modal Data Fusion for Person Authentication using SVM

- **Event:** IEEE Intl. Conference on Acoustic, Speech, and Signal Processing, Phoenix, Arizona, USA, March 15-19, 1999

  **Speaker:** Hervé Bourlard

- **Event:** Workshop on Robust Speech Recognition, Tampere (Finland), May 25-26, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Invited Keynote on "Non-Stationary Multi-Channel (Multi-Stream) Processing Towards Robust and Adaptive ASR"

- **Event:** IEEE Workshop on Neural Networks Signal Processing, Madison, WA, USA, August 23-25, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Member of the Technical Committe and Chairman of the session on Applications

- **Event:** Eurospeech 1999, Budapest, September 6-10, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Chairman of two sessions on "Multi-stream automatic speech recognition" and "Speech enhancement"

- **Event:** Forum Art et Sciences, Institut Universitaire Kurt Bosh, Sion, October 8, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Pattern Recognition

- **Event:** ILASH Seminar Day, Sheffield University, UK, December 3, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Towards Speech Recognition using Speech Models

- **Event:** IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, USA, December 12-15, 1999

  **Speaker:** Hervé Bourlard

  **Title:** Iterative Posterior-Based Keyword Spotting Without Filler Models

- **Event:** Eurospeech 1999, Budapest, September 6-10, 1999

  **Speaker:** Hervé Glotin

  **Title:** A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition

- **Event:** Computational Auditory Scene Analysis (CASA) workshop, Stockholm, July 1999

  **Speaker:** Hervé Glotin

  **Title:** A measure of speech and pitch reliability from voicing

- **Event:** Eurospeech 1999, Budapest, September 6-10, 1999

  **Speaker:** A. Hagen

  **Title:** The full combination sub-bands approach to noise robust HMM/ANN based ASR

- **Event:** European Geophysical Society Congress, The Haag, April 1999

  **Speaker:** M. Kanevski

  **Title:** Multivariate spatial data analysis and modelling with combination of artificial neural networks and geostatistics

- **Event:** European Geophysical Society Congress, The Haag, April 1999

  **Speaker:** M. Kanevski

  **Title:** Advanced learning algorithms for spatial data

- **Event:** STATGIS 1999 International Conference on Geostatisitcs and Geographical Information Systems, Klagengurt, Austria, September 1999

  **Speaker:** M. Kanevski

  **Title:** Geostat Office solution for Geostat+GIS

- **Event:** Workshop on SVM, Geostatistics and Software Development, Lausanne, May-June 1999

  **Speaker:** M. Kanevski

  **Title:** Application of SVM for spatial data and software development

- **Event:** Workshop with RHUIL on SVM, Geostatistics and Data Analysis, Lausanne, November 1999

  **Speaker:** M. Kanevski

  **Title:** Spatial data analysis with Geostat Office

- **Event:** Eurospeech 1999, Budapest, September 6-10, 1999

  **Speaker:** Christopher Kermorvant

  **Title:** A comparison of two strategies for ASR in additive noise: Missing data and spectral substraction

- **Event:** Eurospeech 1999, Budapest, September 6-10, 1999

  **Speaker:** Sacha Krstulović

  **Title:** Extraction of Articulators in X-Ray Image Sequences

- **Event:** ESCA ETRW Workshop on Accessing Information in Spoken Audio, Cambridge, UK, April 19-20, 1999

  **Speaker:** Mikko Kurimo

  **Title:** Latent Semantic Indexing by Self-Organizing Map

- **Event:** Workshop on Self-Organizing Maps, Espoo, Finland, July 1-3, 1999

  **Speaker:** Mikko Kurimo

  **Title:** Indexing Audio Documents by using Latent Semantic Analysis and SOM

- **Event:** International Conference on Computer Analysis of Images and Patterns, Ljubliana, Slovenia, September 1-3, 1999

  **Speaker:** Jürgen Lüttin

  **Title:** Evaluating the Complexity of Databases for Person Identification and Verification

  **Title:** Tracking Articulators in X-ray Movies of the Vocal Tract

- **Event:** IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, USA, June 23-25, 1999

  **Speaker:** Jürgen Lüttin

  **Title:** Audio-Visual Person Verification

- **Event:** Second International Conference on Audio-Visual Biometric Person Authentication, Washington D.C., USA, March 22-23, 1999

  **Speaker:** Jürgen Lüttin

  **Title:** Fast Face Detection using MLP and FFT

- **Event:** MANTRA Seminar, EPFL, Lausanne, May 21, 1999

  **Speaker:** Perry Moerland

  **Title:** DynaBoost: Combining Boosted Hypotheses in a Dynamic Way

- **Event:** International Conference on Artificial Neural Networks, Edinburgh, United Kingdom, September 7-10, 1999

  **Speaker:** Perry Moerland

  **Title:** A comparison of mixture models for density estimation

- **Event:** International Conference on Artificial Neural Networks, Edinburgh, United Kingdom, September 7-10, 1999

  **Speaker:** Perry Moerland

  **Title:** Classification using localized mixtures of experts

- **Event:** ICML'99 Workshop: From Machine Learning to Knowledge Discovery in Databases, Bled, Slovenia, June 27-30, 1999

  **Speaker:** Miguel Moreira

  **Title:** Data binarization by discriminant elimination

- **Event:** Principles of Data Mining and Knowledge Discovery - PKDD 1999, Prague, Czech Republic, September 15-18, 1999

  **Speaker:** Miguel Moreira

  **Title:** Combinatorial Approach for Data Binarization

- **Event:** International Robust'99 Workshop on "Robust Methods for Speech Recognition in Adverse Conditions", Tampere, Finland, May 25-26, 1999

  **Speaker:** Andrew Morris

  **Title:** Different Weighting Schemes in the Full Combination Subbands Approach for Noise Robust ASR

- **Event:** Poster Presentation at the Rencontres Jeunes Chercheurs en Parole, Avignon, November 18-19, 1999

  **Speaker:** Katrin Weber

  **Title:** Une nouvelle approche pour exploiter des dependances avec les HMM

- **Event:** Automatic Speech Recognition and Understanding Workshop Keystone, USA, December 12-15, 1999

  **Speaker:** Katrin Weber

  **Title:** Towards introducing long-term statistics in MUSE for robust speech recognition (Poster of C. Kermorvant and C. Mokbel)

- **Event:** International Robust'99 Workshop on "Robust Methods for Speech Recognition in Adverse Conditions", Tampere, Finland, May 25-26, 1999

  **Speaker:** Katrin Weber and Chafic Mokbel

  **Title:** Missing Feature Theory and Parallel Model Combination for Robust Speech Recogntion (Poster of P. Renevey and A. Drygajlo of EPFL)

# 7 Publications (1998 and 1999)

## 7.1 Books and Book Chapters

[1] F. Beaufays, H. Bourlard, H. Franco, and N. Morgan, *Neural networks in automatic speech recognition*, in to be published in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., Bradford Books, The MIT Press, 2000.

[2] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich, *Traitement de la Parole*, Presses Polytechniques Universitaires Romandes, 2000.

[3] H. Bourlard and N. Morgan, *Connectionist techniques*, in Survey of the State of the Art in Human Language Technology, R. C. et al., ed., Cambridge University Press, 1998, pp. 356–361.

[4] H. Bourlard and N. Morgan, *Hybrid HMM/ANN systems for speech recognition: Overview and new research directions*, in Adaptive Processing of Sequences and Data Structures, C. L. Giles and M. Gori, eds., Lecture Notes in Artificial Intelligence (1387), Springer Verlag, 1998, pp. 389–417.

[5] M. Kurimo, *Indexing audio documents by using latent semantic analysis and som*, in Kohonen Maps, E. Oja and S. Kaski, eds., Elsevier, 1999, pp. 363–374.

[6] J. Luettin, *Speech reading*, in Modern Interface Technology: The Leading Edge, J. Noyes and M. Cooke, eds., Research Studies Press Ltd., 1999, pp. 97–121.

[7] N. Morgan, H. Bourlard, and H. Hermansky, *Automatic speech recognition: an auditory perspective*, in Speech Processing in the Auditory System, S. Greenberg, W. Ainsworth, A. Popper, and R. Fay, eds., Springer Verlag, New York, 2000.

## 7.2 Articles in International Journals

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, *Fusion of face and speech data for person identity verification*, IEEE Transactions on Neural Networks, 10 (1999), pp. 1065–1074.

[2] S. Bengio and Y. Bengio, *Taking on the curse of dimensionality in joint distributions using neural networks*, to appear in IEEE Transaction on Neural Networks special issue on data mining and knowledge discovery, (2000).

[3] E. Mayoraz, *On the complexity of recognizing regions computable by two-layered perceptrons*, Annals Mathematics and Artificial Intelligence, (1999).

[4] P. D. Moerland, E. Fiesler, and I. Saxena, *Discrete all-positive multilayer perceptrons for optical implementation*, Optical Engineering, 37 (1998), pp. 1305–1315.

[5] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, *Multi-stream adaptive evidence combination for noise robust asr*, Speech Communication, (2000).

[6] B. Nedic, F. Bimbot, R. Blouet, J.-F. Bonastre, G. Caloz, J. Cernocky, G. Chollet, G. Durou, C. Fredouille, D. Genoud, G. Gravier, J. Hennebert, J. Kharroubi, I. Magrin-Chagnolleau, T. Merlin, C. Mokbel, D. Petrovska, S. Pigeon, M. Seck, P. Verlinde, and M. Zouhal, *The ELISA systems for the NIST'99 evaluation in speaker detection and tracking*, DSP Journal (Special Issue on the Nist Speaker Recognition Workshop), (1999).

## 7.3 Articles in Conference Proceedings

[1] S. Ben-Yacoub, *Multi-modal data fusion for person authentication using SVM*, in Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99), 1999, pp. 25–30.

[2] S. Ben-Yacoub, B. Fasel, and J. Luettin, *Fast face detection using MLP and FFT*, in Proc. Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), 1999, pp. 31–36.

[3] S. Ben-Yacoub, J.Luettin, K. Jonsson, J. Matas, and J. Kittler, *Audio-visual person verification*, in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 1999, Fort Collins, USA, 1999.

[4] G. Bernardis and H. Bourlard, *Confidence measures in hybrid HMM/ANN speech recognition*, in Proceedings of Workshop on Text, Speech and Dialog (TSD'98) Brno, Czech Republic, September 1998, pp. 159–164.

[5] G. Bernardis and H. Bourlard, *Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems*, in Proceedings of International Conference on Spoken Language Processing (ICSLP'98) Sydney, Australia, 1998, pp. 775–778.

[6] F. Berthommier and H. Glotin, *A measure of speech and pitch reliability from voicing*, in Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), F. Klassner, ed., Computational Auditory Scene Analysis (CASA) workshop, Stockholm, July 1999, Scandinavian AI Society, pp. 61–70.

[7] F. Berthommier and H. Glotin, *A new snr-feature mapping for robust multistream speech recognition*, in Proc. Int. Congress on Phonetic Sciences (ICPhS), B. University Of California, ed., vol. 1 of XIV, San Francisco, August 1999, pp. 711–715.

[8] F. Berthommier, H. Glotin, E. Tessier, and H. Bourlard, *Interfacing of CASA and partial recognition based on a multistream technique*, in ICSLP'98, vol. 4, Sidney, 1998, pp. 1415–1419.

[9] L. Besacier, J. Luettin, G. Maitre, and E. Meurville, *Experimental evaluation of text-dependent speaker verification on laboratory and field test databases in the M2VTS project*, in Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 751–754.

[10] F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Marietho z, C. Mokbel, and H. Mokbel, *An overview of the PICASSO project research activities in speaker verification for telephone applications*, in 6th european conference on speech communication and technology — eurospeech'99, vol. 5, Budapest, Hungary, September 5–10 1999, pp. 1963–1966.

[11] F. Bimbot, H. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J. Pierrot, *An overview of the cave project research activities in speaker verification*, in Reconnaissance du locuteur et ses applications commerciales et criminalistiques, 1998.

[12] H. Bourlard, *Connectionist speech recognition*, in Proceedings of IK'98, Interdisziplinäres Kolleg, Spring Scholl, Günne am Möhnessee, Germany, March 7–14, 1998, pp. 61–89.

[13] H. Bourlard, *Non-stationary multi-channel (multi-stream) processing towards robust and adaptive asr*, in Proc. of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions, 1999.

[14] G. Caloz, C. Jaboulet, J. Mariéthoz, A. Glaeser, and D. Genoud, *Voice-b system*, in IEEE 4th Workshop on Intercative Voice Technology for Telecommunications Applications (IVTTA'98) September 29–30, Torino, Italy, 1998, pp. 107–111.

[15] S. Choi, Y. Lyu, F. Berthommier, H. Glotin, and A. Cichocki, *Blind separation of delayed and superimposed acoustic sources : learning algorithms an experimental study*, in Proc. IEEE Int. Conference on Speech Processing (ICSP), Seoul, September 1999, IEEE.

[16] V. Demyanov, N. Gilardi, M. Kanevski, M. Maignan, and V. Polishchuk, *Decision-oriented environmental mapping with radial basis function neural networks*, in Intelligent techniques for Spatio-Temporal Data Analysis in Environmental Applications. Workshop W07, 1999, pp. 33–42.

[17] V. Demyanov, M. Kanevski, M. Maignan, E. Savelieva, V. Timonin, S. Chernov, and G. Piller, *Indoor radon risk assessment with geostatistics and artificial neural networks*, in Geostatistical congress 2000, 2000.

[18] V. Demyanov, M. Kanevski, E. Savelieva, V. Timonin, and S. Chernov, *Neural network residual stochastic co-simulation for environmental data analysis*, in Neural Computation 2000, 2000.

[19] S. Dupont and J. Luettin, *Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database*, in Proc. 5th Int. Conf. on Spoken Language Processing, vol. 4, 1998, pp. 1283–1286.

[20] B. Frédéric and G. Hervé, *A CASA front-end using the harmonicity cue for speech enhancement in loud noise*, in submited in Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Istambul, 2000, IEEE.

[21] C. Fredouille, J. Mariéthoz, C. Jaboulet, J. Hennebert, C. Mokbel, and F. Bimbot, *Behavior of a bayesian adaptation method for incremental enrollment in speaker verification*, in ICASSP2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5–9 2000.

[22] D. Genoud and G. Chollet, *Speech pre-processing against intentional imposture in speaker recognition*, in Proceedings of ICSLP, Sidney, 1998.

[23] D. Genoud and G. Chollet, *Voice transformation, a tool for imposture of speaker verification*, in Proceedings of International Phonetic Science conference IPS98, Washington, 1998.

[24] D. Genoud and G. Chollet, *Deliberate imposture: a challenge for automatic speaker verification systems*, in Proceedings of the European Conference on Speech Communication and Technology, 1999.

[25] D. Genoud, M. Moreira, and E. Mayoraz, *Text dependent speaker verification using binary classifiers*, in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing — ICASSP'98, vol. I, IEEE, IEEE, May 1998, pp. 129–132.

[26] N. Gilardi, M. Kanevski, M. Maignan, and E. Mayoraz, *Environmental and pollution spatial data classification with support vector machines and geostatistics*, in Intelligent techniques for Spatio-Temporal Data Analysis in Environmental Applications. Workshop W07, 1999, pp. 43–51.

[27] N. Gilardi, M. Kanevski, M. Maignan, and E. Mayoraz, *Environmental and pollution spatial data classification with support vector machines and geostatistics*, in Geostatistical congress 2000, 2000.

[28] H. Glotin, F. Berthommier, and E. Tessier, *A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition*, in Proc. European Conf. on Speech Communication and Technology (EUROSPEECH), vol. 5, september 1999, pp. 2351–2354.

[29] H. Glotin, E. Tessier, H. Bourlard, and F. Berthommier, *Reconnaissance multi-bandes de la parole bruitée par couplage entre les niveaux primitifs et d'identification*, in Journées Etude Parole - Martigny, Juin 1998.

[30] H. Glotin, E. Tessier, H. Bourlard, and F. Berthommier, *Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes*, in Neurosciences et Sciences de l'Ingénieur'98 - Munster, CNRS, Mai 1998.

[31] A. Hagen, A. Morris, and H. Bourlard, *Different weighting schemes in the full combination subbands approach for noise robust asr*, in Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, May 1999.

[32] H.Glotin, F. Berthommier, E. Tessier, and H. Bourlard, *Interfacing of CASA and multistream recognition*, in TSD'98-Text, Speech and Dialog International Workshop, BRNO-Czech Republic, Sept 1998.

[33] C. Kermorvant and C. Mokbel, *Towards introducing long-term statistics in muse for robust speech recognition*, in Automatic Speech Recognition and Understanding (ASRU) workshop, Keystone, Colorado, USA, December 1999.

[34] C. Kermorvant and A. Morris, *A comparison of two strategies for asr in additive noise : Missing data and spectral subtraction*, in 6th European Conference on Speech Communication and Technology — Eurospeech'99, Budapest, Hungary, September, 5–10 1999.

[35] S. Krstulović, *LPC-based inversion of the DRM articulatory model*, in Proc. Eurospeech'99, 1999.

[36] S. Krstulović, *LPC modeling with speech production constraints*, in Proc. 5th Speech Production Seminar (to appear), 2000.

[37] M. Kurimo and C. Mokbel, *Latent semantic indexing by self-organizing map*, in ESCA ETRW workshop on Accessing Information in Spoken Audio, Cambridge, UK, April 1999, pp. 25–30.

[38] J. Luettin and S. Ben-Yacoub, *Robust person verification based on speech and facial images*, in Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 991–994.

[39] J. Luettin and S. Dupont, *Continuous audio-visual speech recognition*, in Proc. 5th European Conference on Computer Vision, vol. II of Lecture Notes in Computer Science, Springer Verlag, 1998, pp. 657–673.

[40] J. Mariéthoz, D. Genoud, F. Bimbot, and C. Mokbel, *Client / world model synchronous alignement for speaker verification*, in 6th European Conference on Speech Communication and Technology — Eurospeech'99, Budapest, Hungary, September 5–10 1999.

[41] E. Mayoraz and M. Moreira, *Combinatorial approach for data binarization*, in Principles of Data Mining and Knowledge Discovery: third european conference; proceedings / PKDD'99, J. Zytkow and J. Rauch, eds., vol. 1704 of Lecture Notes in Artificial Intelligence, Springer, 1999, pp. 442–447.

[42] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, *XM2VTSDB: The extended M2VTS database*, in Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99), 1999.

[43] P. Moerland, *Classification using localized mixtures of experts*, in Proceedings of the International Conference on Artificial Neural Networks (ICANN'99), vol. 2, London: IEE, 1999, pp. 838–843.

[44] P. Moerland, *A comparison of mixture models for density estimation*, in Proceedings of the International Conference on Artificial Neural Networks (ICANN'99), vol. 1, London: IEE, 1999, pp. 25–30.

[45] C. Mokbel and O. Collin, *Incremental enrollment of speech recognizers*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99), Phoenix, Arizona, USA, 1999.

[46] M. Moreira, A. Hertz, and E. Mayoraz, *Data binarization by discriminant elimination*, in Proceedings of the ICML-99 Workshop: From Machine Learning to Knowledge Discovery in Databases, I. Bruha and M. Bohanec, eds., 1999, pp. 51–60.

[47] M. Moreira and E. Mayoraz, *Improved pairwise coupling classification with correcting classifiers*, in Machine Learning: ECML-98, C. Nédellec and C. Rouveirol, eds., vol. 1398 of Lecture Notes in Artificial Intelligence, Springer, April 1998, pp. 160–171.

[48] A. Morris, A. Hagen, and H. Bourlard, *The full combination sub-bands approach to noise robust hmm/ann based asr*, in 6th European Conference on Speech Communication and Technology — Eurospeech'99, Budapest, Hungary, September 5–10 1999.

[49] B. Nedic, G. Gravier, J. Kharroubi, G. Chollet, D. Petrovska, G. Durou, F. Bimbot, R. Blouet, M. Seck, J.-F. Bonastre, C. Fredouille, T. Merlin, I. Magrin-Chagnolleau, S. Pigeon, P. Verlinde, and J. Cernocky, *The Elisa'99 speaker recognition and tracking systems*, in IEEE Workshop on Automatic Advanced Technologies, 1999.

[50] D. Petrovska, J. Hennebert, H. Melin, and D. Genoud, *Polycost: a telephone-speech database for speaker recognition*, in Reconnaissance du locuteur et ses applications commerciales et criminalistiques, 1998.

[51] J. Pierrot, J. Lindberg, J. Koolwaaij, H. Hutter, D. Genoud, M. Blomberg, and F. Bimbot, *A comparison of a priori threshold setting procedures for speaker verification in the CAVE project*, in ICASSP 98, 1998.

[52] V. Polishchuk and M. Kanevski, *Comparison of unsupervised and supervised training of rbf neural networks. case study: Mapping of contamination data*, in Neural Computation 2000, 2000.

[53] G. Richard, Y. Menguy, I. Guis, N. Suaudeau, J. Boudy, P. Lockwood, C. Fernndez, F. Fernndez, D. Garcia-Plaza, C. Kotropoulos, A. Tefas, I. Pitas, R. Heimgartner, P. Ryser, C. Beumier, P. Verlinde, S. Pigeon, G. Matas, J. Kittler, J. Bigün, Y. Abdeljaoued, E. Meurville, L. Besacier, M. Ansorge, G. Maitre, J. Luettin, S. Ben-Yacoub, B. Ruiz, J. Cortés, and K. Aldama, *Multi modal verification for teleservices and security applications*, in IEEE International Conference on Multimedia Computing and Systems, 1999.

[54] M.-C. Silaghi and H. Bourlard, *Iterative posterior-based keyword spotting without filler models*, in Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU'99) Workshop, 1999.

[55] M.-C. Silaghi and H. Bourlard, *Iterative posterior-based keyword spotting without filler models*, in Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[56] E. Tessier, F. Berthommier, H. Glotin, and S. Choi, *A CASA front-end using the localisation cue for segregation and then cocktail-party speech recognition*, in Proc. IEEE Int. Conference on Speech Processing (ICSP), Seoul, September 1999, IEEE.

[57] G. Thimm, *Tracking articulators in x-ray movies of the vocal tract*, in 8th Int. Conf. Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, Springer Verlag, 1999, pp. 126–133.

[58] G. Thimm, S. Ben-Yacoub, and J. Luettin, *Evaluating the complexity of databases for person identification and verification*, in 8th Int. Conf. Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, Springer Verlag, 1999, pp. 49–56.

[59] G. Thimm and J. Luettin, *Illumination-robust pattern matching using distorted color histograms*, in Lecture Notes in Computer Science (5th Open German-Russian Workshop on Pattern Recognition and Image Understanding), Springer Verlag, September 21 - 25, 1998.

[60] G. Thimm and J. Luettin, *Extraction of articulators in x-ray image sequences*, in Proceedings of the European Conference on Speech Communication and Technology, 1999, pp. 157–160.

[61] P. Verlinde, G. Maître, and E. Mayoraz, *Decision fusion using a multi-linear classifier*, in 1st International Conference on Multisource-Multisensor Data Fusion, July 1998.

## 7.4 IDIAP Research Reports

[1] E. Alpaydin, *Combined 5x2cv f-test for comparing supervised classification learning algorithms*, IDIAP-RR 4, IDIAP, 1998.

[2] E. Alpaydin and E. Mayoraz, *Combining linear dichomotizers to construct nonlinear polychotomizers*, IDIAP-RR 5, IDIAP, 1998.

[3] G. Bernardis, H. Bourlard, M. Rajman, and J.-C. Chappelier, *INtegrating SPEech acoustic and linguistic Constraints: Baseline System Development*, IDIAP-RR 21, IDIAP, 1999.

[4] F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariethoz, C. Mokbel, and H. Mokbel, *An overview of the PICASSO project research activities in speaker verification for telephone applications*, IDIAP-RR 24, IDIAP, 1999.

[5] H. Bourlard, *Introduction à la reconnaissance de la parole et du locuteur*, IDIAP-RR 13, IDIAP, 1998.

[6] H. Bourlard and N. Morgan, *Speaker Verification: A Quick Overview*, IDIAP-RR 12, IDIAP, 1998.

[7] B. Fasel and J. Luettin, *Facial expression analysis and recognition: A survey*, IDIAP-RR 19, IDIAP, 1999.

[8] B. Fasel and J. Luettin, *Recognition of asymmetric facial action unit activities and intensities*, IDIAP-RR 22, IDIAP, 1999.

[9] C. Fredouille, J. Mariéthoz, C. Jaboulet, J. Hennebert, C. Mokbel, and F. Bimbot, *Behavior of a bayesian adaptation method for incremental enrollment in speaker verification*, IDIAP-RR 02, IDIAP, 2000.

[10] A. Hagen, A. Morris, and H. Bourlard, *Subband-based speech recognition in noisy conditions: The full combination approach*, IDIAP-RR 15, IDIAP, 1998.

[11] M. Kanevski and N. Gilardi, *Numerical experiments with support vector machines*, IDIAP-RR 15, IDIAP, 1999.

[12] M. Kanevski, N. Gilardi, E. Mayoraz, and M. Maignan, *Environmental spatial data classification with support vector machines*, IDIAP-RR 7, IDIAP, 1999.

[13] K. Keller, S. Ben-Yacoub, and C. Mokbel, *Combining wavelet-domain hidden markov trees with hidden markov models*, IDIAP-RR 14, IDIAP, 1999.

[14] C. Kermorvant, *A comparison of noise reduction techniques for robust speech recognition*, IDIAP-RR 10, IDIAP, 1999.

[15] S. Krstulović, *Acoustico-articulatory inversion of the DRM model through inverse filtering*, IDIAP-RR 16, IDIAP, 1998.

[16] M. Kurimo, *Fast latent semantic indexing of spoken documents by using self-organizing maps*, IDIAP-RR 20, IDIAP, 1999.

[17] J. Luettin, *Speaker verification experiments on the XM2VTS database*, IDIAP-RR 2, IDIAP, 1999.

[18] J. Mariéthoz, D. Genoud, F. Bimbot, and C. Mokbel, *Client / world model synchronous alignement for speaker verification*, IDIAP-RR 23, IDIAP, 1999.

[19] J. Mariéthoz and C. Mokbel, *Synchronous alignment*, IDIAP-RR 06, IDIAP, 1999.

[20] E. Mayoraz and E. Alpaydin, *Support vector machine for multiclass classification*, IDIAP-RR 6, IDIAP, 1998.

[21] P. Moerland, *Localized mixtures of experts*, IDIAP-RR 14, IDIAP, 1998.

[22] P. Moerland and E. Mayoraz, *Dynaboost: Combining boosted hypotheses in a dynamic way*, IDIAP-RR 9, IDIAP, 1999.

[23] M.-C. Silaghi and H. Bourlard, *Iterative posterior-based keyword spotting without filler models: Iterative viterbi decoding and one-pass approach*, IDIAP-RR 27, IDIAP, 1999.

[24] G. Thimm, *Segmentation of X-ray image sequences showing the vocal tract*, IDIAP-RR 1, IDIAP, January 1999.

[25] G. Thimm, *Segmentation of X-ray image sequences showing the vocal tract (with tool documentation)*, IDIAP-RR 1, IDIAP, January 1999.

[26] G. Thimm, S. Ben-Yacoub, and J. Luettin, *Evaluating the complexity of databases for person identification and verification*, IDIAP-RR 10, IDIAP, August 1998.

[27] G. Thimm and J. Luettin, *Illumination-robust pattern matching using distorted color histograms*, IDIAP-RR 9, IDIAP, June 1998.

[28] G. Thimm and J. Luettin, *Optimal parameterization of point distribution models*, IDIAP-RR 01, IDIAP, 1998.

[29] A. Vinciarelli and J. Luettin, *Off-line cursive script recognition based on continuous density HMM*, IDIAP-RR 25, IDIAP, 1999.

## 7.5 IDIAP Communications

[1] J. M. Andersen, *Baseline system for hybrid speech recognition on french (experiments on bref)*, IDIAP-COM 07, IDIAP, 1998.

[2] B. Fasel, *Fast multi-scale face detection*, IDIAP-COM 04, IDIAP, 1998.

[3] J. Luettin and G. Maître, *Evaluation protocol for the extended M2VTS database (XM2VTSDB)*, IDIAP-COM 05, IDIAP, 1998.

[4] A. Morris, *Latent variable decomposition for posteriors or likelihood based subband asr*, IDIAP-COM 04, IDIAP, 1999.

## 7.6 Other Documents

[1] D. Genoud, *Reconnaissance et Transformation de Locuteurs*, PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, January 1999.