IDIAP COMMUNICATION

ACTIVITY REPORT 2003

IDIAP–COM 2004-01

FEBRUARY 2004

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

Est. 1991

INTERACTIVE MULTIMODAL INFORMATION MANAGEMENT

Est. 2001

IDIAP Research Institute

Contents

1 Introduction (in English)

Having celebrated its 10th anniversary in 2001 and the granting of the NCCR IM2, and after 2002 was marked by very strong growth (symbolized by the construction of a second building) the IDIAP Research Institute has seen a year of consolidation, in term of personnel, infrastructure and finance in 2003. However, this relative calm did not prevent IDIAP members from achieving a remarkable number of successes, which include important new projects, prestigious awards, high numbers of publications, 4 PhD theses, one spin-off and, last but not least, several world-class researchers joining the Institute.

Since it was established in 1991 by the City of Martigny, the Canton of Valais, the Swiss Federal Institute of Technology in Lausanne, the University of Geneva and Swisscom, IDIAP has remained faithful to its original missions: Fundamental **research** at the highest level in its areas of expertise, thus ensuring a leading role nation-, european-, and world-wide; **education** of young researchers through the hiring of PhD students but also by attracting talented undergraduates to allow them to discover the world of research; **transfer** of knowledge and technologies to guarantee the best possible dissemination of research results in the scientific world but also within tight collaborations with industry.

To fulfill these complementary missions, IDIAP relies on various sources of funding. The largest part – more than 70% – comes from competitive research grants based on projects and funded by the Swiss National Science Foundation (10% directly and 33% through the NCCR), through the participation in European projects (funded so far by the Swiss Federal Office for Education and Science (OFES), 16%), through one US-DARPA (*Defense Advanced Research Projects Agency*) project (7%). Industry-oriented projects have also been funded by the Commission for Technology and Innovation (CTI, 3%) and by private companies (2%). Aside from this research funding, IDIAP gratefully acknowledges support from the City of Martigny (10%), the Canton of Valais (8%), the Swiss Confederation (5%) and the Loterie Romande (6%), which enables the Institute to provide optimal administrative support and excellent infrastructure.

If 2003 was the year of consolidation, it was also an opportunity to rethink the structures of the Institute and to adapt them to strong growth. On the scientific side, the three research groups will leave room to a flatter structure featuring 6 thematic areas (machine learning, speech processing, computer vision, media indexing, biometric authentication, multimodal interaction). Overlaps between these areas are important, which also facilitate the exchange of information among researchers. The structure of this activity report already reflects this evolution as section 5 is more elaborate than previously. Structures governing the Institute have also been revised, and 2004 will see the replacement of the Foundation, Direction and Scientific Councils with a much more focused Foundation Council and Scientific Council, as well as a new Advisory Board more in line with the new worldwide dimension of IDIAP.

The year 2003 also saw a stronger anchoring of IDIAP in the swiss academic landscape through a better collaboration with EPFL. This collaboration was settled on June, 2nd 2003 with the signature of a collaboration agreement which fixes the status of most of the IDIAP PhD students as well as the academic activity of the senior researchers. This agreement will become fully effective during the coming year.

Some figures to complete this introduction. It is gratifying to see that the massive increase of the scientific personnel in 2002 (+15 people, or +25%) is reflected by an equally massive increase of the publications over the last two years, with 24 journal papers (16 in 2001/2002), 91 conference papers (59), and 79 internal or submitted reports (54).

Many other successes have marked the year 2003 at IDIAP, from the IM2 NCCR to the new European Integrated Project AMI, from the Spiderphone spin-off to the integration in strategy of the Canton du Valais for economic development. These are recalled and explained in section 4.

The management and personnel of IDIAP wishes you a pleasant reading of this annual report and redirects you to the web site www.idiap.ch for the latest information. On our side, we are working hard to make 2004 even better.

2 Introduction (en français)

Après avoir célébré son 10e anniversaire en 2001 avec l'attribution du PRN IM2, après une année 2002 marquée par une très forte croissance et symbolisée par la construction d'un second bâtiment, l'Institut de Recherche IDIAP de Martigny a vécu en 2003 une année de consolidation, aussi bien sur le plan de son personnel, de ses infrastructures que de ses finances. Ce calme relatif n'a cependant pas empêché les membres de l'IDIAP de décrocher un grand nombre de succès, au nombre desquels on trouve d'importants nouveaux projets, des prix prestigieux, un nombre élevé de publications, 4 thèses de doctorat, une spin-off, enfin et surtout, plusieurs chercheurs de renommée mondiale qui sont venus renforcer notre effectifs.

Depuis sa fondation en 1991 par la Ville de Martigny, le Canton du Valais, l'Ecole Polytechnique Fédérale de Lausanne, l'Université de Genève et Swisscom, l'IDIAP est resté fidèle aux missions qui lui ont été confiées : La **recherche** fondamentale de très haut niveau dans les domaines de compétence de l'Institut, qui doit lui assurer une place de leader au plan national, européen et mondial; la **formation** de jeunes chercheurs au travers des doctorants mais aussi en attirant des universitaires de talent avant la fin de leurs études pour leur donner l'occasion de découvrir le monde de la recherche; la **valorisation** des résultats obtenus par le biais d'une diffusion aussi large que possible dans les milieux scientifiques, mais aussi et surtout par des contacts étroits avec les milieux industriels pour assurer le transfert des connaissances et des technologies.

Pour remplir ces missions complémentaires, l'IDIAP peut compter sur diverses sources de financement. La part de loin la plus importante – plus de 70% – provient de projets de recherche attribués sur une base compétitive par le Fonds National Suisse pour la Recherche Scientifique (10% directement et 33% via le PRN), via la participation à des projets européens (financé jusqu'à maintenant par l'Office Fédéral de l'Education et de la Science (OFES), 16%), via un projet financé par le DARPA (*Defense Advanced Research Projects Agency*) américain (7%). Toujours dans la catégorie des projets décrochés par l'IDIAP, mais en relation avec le monde industriel, on trouve encore un apport de la Commission pour la Technologie et l'Innovation (CTI, 3%) ainsi que des contributions d'entreprises privées (2%). Pour compléter ces moyens qui financent avant tout le personel scientifique, l'IDIAP peut fort heureusement compter sur un soutien de la Ville de Martigny (10%), du Canton du Valais (8%), de la Confédération (5%) et de la Loterie Romande (6%), soutien qui lui permet d'offrir un appui administratif et structurel optimal.

Si 2003 a été une année de consolidation, elle a aussi été l'occasion de repenser les structures de l'Institut et de les adapter à la forte croissance. Sur le plan scientifique, les trois groupes de recherche qui composaient l'Institut s'estompent pour faire place à une structure plus plate caractérisée par 6 directions thématiques (machine learning, speech processing, computer vision, media indexing, biometric authentication, multimodal interaction). Les recouvrements entre ces thèmes sont beaucoup plus important qu'auparavant, ce qui garanti également une meilleure circulation de l'information parmi chercheurs. La structure de ce rapport d'activité reflète déjà cette évolution puisque la section 5 est bien plus étoffé que par le passé. Les structures chapeautant l'Institut ont également fait l'objet d'une révision et 2004 verra le remplacement des actuels Conseils de Fondation, de Direction et Scientifique, par un Conseil de Fondation et un Conseil Scientifique plus ciblés, ainsi que par un nouveau Comité Stratégique (Advisory Board) plus en ligne avec la dimension mondiale qu'a atteint l'IDIAP.

2003 a également vu un ancrage plus important de l'IDIAP dans le paysage académique suisse au travers d'une collaboration plus étroite avec l'EPFL. Cette collaboration s'est concrétisée le 2 juin 2004 par la signature d'une convention de collaboration qui ancre dans les faits le statut de la majorité des doctorants de l'IDIAP ainsi que l'activité académique des chercheurs seniors. Cette convention développera tous ses effets au cours de l'année à venir.

Quelques chiffres pour terminer cette introduction. Il est réjouissant de constater que l'augmentation massive du personnel scientifique en 2002 (+15 personnes, ou +25%) se traduit par une augmentation tout autant importante des publications sur les deux dernières années, avec 24 articles dans des journaux (16 en 2001/2002), 91 articles dans des conférences (59) et 79 rapports internes ou en cours de publications (54).

De nombreux autres succès ont jalonné l'année 2003 à l'IDIAP, du PRN IM2 au nouveau projet européen AMI, de la spin-off Spiderphone à l'intégration à la politique du développement économique du Canton du Valais. Ils sont repris et développés dans la section 4.

La direction et le personnel de l'IDIAP vous souhaite une agréable lecture de ce rapport annuel tout en vous renvoyant au site `www.idiap.ch` pour les informations les plus récentes. De notre côté, nous travaillons activement à faire encore mieux en 2004.

3 Staff

General contact information:

Mail: Institut de Recherche IDIAP
Rue du Simplon 4, CP 592
CH–1920 Martigny (VS)
Switzerland

Phone: +41 - 27 - 721 77 11

Fax: +41 - 27 - 721 77 12

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

3.1 Scientific Staff

Mr	Jitendra AJMERA Jitendra.Ajmera@idiap.ch	Research Assistant +41 27 721 77 48	
Mr	Silèye BA Sileye.Ba@idiap.ch	Research Assistant +41 27 721 77 61	
Mr	Marc BARNARD Mark.Barnard@idiap.ch	Research Assistant +41 27 721 77 29	
Dr	Samy BENGIO Samy.Bengio@idiap.ch	Senior Research Scientist +41 27 721 77 39	
Mr	Mohamed F. BENZEGHIBA Mohamed.Benzeghiba@idiap.ch	Research Assistant +41 27 721 77 41	
Prof.	Hervé BOURLARD Herve.Bourlard.@idiap.ch	Director +41 27 721 77 20	
Mr	Fabien CARDINAUX Fabien.Cardinaux@idiap.ch	Research Assistant +41 27 721 77 55	
Dr	Datong CHEN Datong.Chen@idiap.ch	Research Assistant	→ 30.11.03
Ms	Silvia CHIAPPA Silvia.Chiappa@idiap.ch	Research Assistant +41 27 721 77 30	
Mr	Ronan COLLOBERT Ronan.Collobert@idiap.ch	Research Assistant +41 27 721 77 31	
Mr	Christos DIMITRAKAKIS Christos.Dimitrakakis@idiap.ch	Research Assistant +41 27 721 77 40	
Dr	John DINES John.Dines@idiap.ch	Research Scientist +41 27 721 77 60	01.03.03 →
Mr	Mike FLYNN Mike.Flynn@idiap.ch	Senior development engineer +41 27 721 77 78	01.06.03 →
Mr	Beat FASEL Beat.Fasel@idiap.ch	Research Assistant	→ 28.02.03

Dr	Daniel GATICA-PEREZ Daniel.Gatica-Perez@idiap.ch	Senior Research Scientist +41 27 721 77 33	
Mr	David GRANGIER David.Grangier@idiap.ch	Research Assistant +41 27 721 77 23	01.10.03 →
Mr	Maël GUILLEMOT Mael.Guillemot@idiap.ch	Development engineer +41 27 721 77 64	
Prof.	Hynek HERMANSKY Hynek.Hermansky@idiap.ch	Senior Research Scientist +41 27 721 77 73	01.07.03 →
Mr	Shajith IKBAL Shajith.Ikbal@idiap.ch	Research Assistant +41 27 721 77 46	
Ms	Agnès JUST Agnes.Just@idiap.ch	Research Assistant +41 27 721 77 68	
Ms	Mikaela KELLER Mikaela.Keller@idiap.ch	Research Assistant +41 27 721 77 75	
Mr	Itshak LAPIDOT Itsak.Lapidot@idiap.ch	Research Scientist	→ 28.02.03
Mr	Guillaume LATHOUD Guillaume.Lathoud@idiap.ch	Research Assistant +41 27 721 77 63	
Mr	Quan LE Quan.Le@idiap.ch	Research Assistant	→ 31.10.03
Mr	Mathew MAGIMAI DOSS Mathew@idiap.ch	Research Assistant +41 27 721 77 51	
Ms	Viktoria MAIER Viktoria.Maier@idiap.ch	Research Assistant +41 27 721 77 59	01.08.03 →
Dr	Sebastien MARCEL Sebastien.Marcel@idiap.ch	Senior Research Scientist +41 27 721 77 27	
Mrs	Christine MARCEL Christine.Marcel@idiap.ch	Development engineer +41 27 721 77 50	
Mr	Johnny MARIÉTHOZ Johnny.Mariethoz@idiap.ch	Development engineer +41 27 721 77 44	
Mr	Olivier MASSON Olivier.Masson@idiap.ch	Development engineer +41 27 721 77 66	
Dr	Iain MCCOWAN Iain.Mccowan@idiap.ch	Senior Research Scientist +41 27 721 77 32	
Mr	Michael MCGREEVY Michael.McGreevy@idiap.ch	Research Assistant +41 27 721 77 35	15.01.03 →
Dr	José MILLAN José.Millan@idiap.ch	Senior Research Scientist +41 27 721 77 70	
Mr	Hemant MISRA Hemant.Misra@idiap.ch	Research Assistant +41 27 721 77 57	

Mr	Florent MONAY Florent.Monay@idiap.ch	Research Assistant +41 27 721 77 69	
Mr	Darren MOORE Darren.Moore@idiap.ch	Development engineer +41 27 721 77 34	
Dr	Jean-Marc ODOBEZ Jean-Marc.Odobez@idiap.ch	Senior Research Scientist +41 27 721 77 26	
Mr	Norman POH HOON THIAN Norman.Poh@idiap.ch	Research Assistant +41 27 721 77 53	
Mr	Alexei POZDNOUKHOV Alexei.Pozdnoukhov@idiap.ch	Research Assistant +41 27 721 77 65	
Mr	Pedro QUELHAS Pedro.Quelhas@idiap.ch	Research Assistant +41 27 721 77 74	
Mr	Yann RODRIGUEZ Yann.Rodriguez@idiap.ch	Research Assistant +41 27 721 77 72	
Dr	Conrad SANDERSON Conrad.Sanderson@idiap.ch	Research Scientist +41 27 721 77 43	
Mr	Sunil SIVADAS Sunil.Sivadas@idiap.ch	Research Assistant +41 27 721 77 79	01.09.03 →
Mr	Kevin SMITH Kevin.Smith@idiap.ch	Research Assistant +41 27 721 77 67	
Dr	Todd STEPHENSON Todd.Stephenson@idiap.ch	Research Assistant	→ 30.06.03
Mr	Vivek TYAGY Vivek.Tyagi@idiap.ch	Research Assistant	→ 31.10.03
Dr	Alessandro VINCIARELLI Alessandro.Vinciarelli@idiap.ch	Research Scientist +41 27 721 77 24	
Dr	Katrin WEBER Katrin.Weber@idiap.ch	Research Assistant	→ 31.05.03
Dr	Pierre WELLNER Pierre.Wellner@idiap.ch	Senior Research Scientist +41 27 721 77 62	
Mr	Dong ZHANG Dong.Zhang@idiap.ch	Research Assistant +41 27 721 77 76	01.08.03 →

3.2 Students

Mr	Aïssa AIT-HASSOU Aissa.Ait-Hassou@idiap.ch	01.02.03 → 31.05.03
Mr	Guillermo Zapata ARADILLA Guillermo.Aradilla@idiap.ch	01.09.03 → 29.02.04
Mr	François BESSARD Francois.Bessard@idiap.ch	01.07.03 → 31.12.03
Mr	Bastien CRETTOL Bastien.Crettol@idiap.ch	01.09.03 → 31.12.03
Mr	Emmanuel GABBUD Emmanuel.Gabbud@idiap.ch	01.03.03 → 31.05.03
Mr	David GRANGIER David.Grangier@idiap.ch	01.03.03 → 30.09.03
Mr	Frédéric KOTTELAT Frederic.Kottelat@idiap.ch	01.01.03 → 30.06.03
Mr	Jérôme KOWALCZYK Jerome.Kowalczyk@idiap.ch	01.12.03 → 30.06.04
Ms	Viktoria MAIER Viktoria.Maier@idiap.ch	01.02.03 → 30.06.03
Mr	Jean-Sébastien SENÉCAL Jean-Sebastien.Senecal@idiap.ch	15.05.03 → 15.09.03
Mr	Julien TIPHAIGNE Julien.Tiphaigne@idiap.ch	01.12.03 → 30.06.04

3.3 System and Development Staff

Mr	Tristan CARRON Tristan.Carron@idiap.ch	System engineer +41 27 721 77 77	01.01.03 →
Mr	Thierry COLLADO Thierry.Collado@idiap.ch	Webmaster & system engineer +41 27 721 77 42	
Mr	Norbert CRETTOL Norbert.Crettol@idiap.ch	System engineer +41 27 721 77 25	
Mr	Frank FORMAZ Frank.Formaz@idiap.ch	System Management Group Leader +41 27 721 77 28	
Mrs	Haiyan WANG Haiyan.Wang@idiap.ch	Development engineer	→ 30.11.03

3.4 Administrative Staff

Dr	Jean-Albert FERREZ Jean-Albert.Ferrez@idiap.ch	Deputy Director +41 27 721 77 19	
Mr	Pierre DAL PONT Pierre.DalPont@idiap.ch	Financial Manager +41 27 721 77 45	
Ms	Nancy-Lara ROBYR Nancy-Lara.Robyr@idiap.ch	Program Manager +41 27 721 77 18	01.08.03 →
Mrs	Sylvie MILLIUS Sylvie.Millius@idiap.ch	Secretary +41 27 721 77 21	
Mrs	Nadine ROUSSEAU Nadine.Rousseau@idiap.ch	Secretary +41 27 721 77 22	
Mrs	Joanne SCHULZ (MOORE) Joanne.Schulz@idiap.ch	HR assistant +41 27 721 77 49	
Ms	Rosanna BARBUTO Rosanna.Barbuto@idiap.ch	Public Relations	→ 30.04.03
Mr	Michel SALAMIN	French teacher	

4 Major events in 2003

The second year of the IM2 NCCR

Now in its second year, the IM2 NCCR still plays a major role in the activities of IDIAP. While 2002 saw the setting up of the research and related activities, 2003 was dominated by the leveraging that IDIAP researchers have been able to accomplish thanks to IM2. The most visible example is of course the EC project AMI (see below), but other projects, workshops, spin-offs also result in some way or have benefited from the NCCR.

As a brief reminder, we list some of the major IM2-related events of 2003, the interested reader will find more information in the annual NCCR progress report, available upon request.

A new Deputy Director

Following Prof. Murat Kunt's resignation, EPFL Professor Roland Siegwart, head of the Institute of Systems Engineering, has been nominated to second Prof. Hervé Bourlard. In particular, Roland will be in charge of coordinating education activities within IM2.

ICSI fellowship

In the framework of the ongoing ICSI-IM2 exchange program, ICSI has opened a number of visitor fellowship positions for the fall of 2003 and for 2004. These positions are open to postdocs and PhD students working within IM2 as well as to outstanding undergraduate students who intend to continue their career with a PhD position within IM2.

The SNSF Review Panel Site Visit

In addition to the annual progress report, the NCCR was assessed during a two day visit of the SNSF-appointed Review Panel. The panel members are: Dr Phil Janson (SNSF, Chairman), Prof. Marco Baggiolini (SNSF), Dr Giordano Bruno Beretta (Hewlett Packard Laboratories, Palo Alto, USA), Prof. Shih-Fu Chang (Columbia University, New York, USA), Prof. Beat Hirsbrunner (SNSF), Prof. Mari Ostendorf (University of Washington, Seattle, USA), Prof. Steve Renals (University of Sheffield, UK), Prof. Gerhard Rigoll (Technische Universität München, D). Prof. Ramesh Jain (Georgia Tech, Atlanta, USA) and Prof. Dan Jurafsky (University of Colorado, Boulder, USA) have joined the panel in 2003, while we regret the resignation of Dr Andrew William Senior.

The panel reports to the SNSF, and the feedback IDIAP received was highly positive.

The Scientific and Industrial Advisory Boards Meeting

The first joint meeting of the IM2 Scientific and Industrial Advisory Boards took place in Martigny on February 13 and 14. The two day meeting started with a general presentation of the NCCR and its objectives, and then focused on three topics: input modalities, multimodal processing, and applications.

The IM2 Summer Institute

On October 6-7-8, 2003, the "Centre de Conférence du Régent" in Crans-Montana hosted the second internal workshop that brought together more than 100 IM2 scientists.

The IM2 web site, `http://www.im2.ch/`

IDIAP hosts a common web site that acts as an entry point for all activities related to the NCCR. The web site was completely re-designed in 2003.

On this web site, one can also find copies of the monthly IM2 Newsletter, featuring activities of the NCCR and related news. A hard copy of this Newsletter can also be received by regular mail upon request to the IDIAP secretaries.

A new FP6 Integrated Project: Augmented Multiparty Interaction (AMI)

One of the big successes of IDIAP in 2003 was the acceptance by the EC of a major "Integrated Project" with IDIAP as initiator and main coordinator. The project, referred to as "Augmented Multi-party Interaction" (AMI), is an Integrated Project from the newly launched FP6-IST programme, benefiting from substantial funding. AMI involves 14 European partners and 1 US partner, involving 4 research institutes, 5 academic partners, 5 industries, and one standard representative (W3C). As with IM2, the project will have to show significant performance in the areas related to research, training and technology transfer. Ranked first (out of 20 high-level proposals) at the end of the selection process, the success of this project proposal is a direct consequence of the success of our projects such as IM2 and M4.

As further discussed in www.amiproject.org, AMI targets computer-enhanced multi-modal interaction in the context of meetings. Directly building upon IM2, AMI however focuses much more on human-to-human communication, i.e., mainly using computers with the goal to *enhance human-human communication* in face-to-face meetings, as well as in remote meetings (not currently addressed in IM2).

Other new FP6 projects

IDIAP is part of several other new FP6 projects:

PASCAL (Pattern analysis, statistical modelling and computational learning) is a Network of Excellence also in the "Multimodal Interfaces" Strategic Objective. The EC made it explicit that they expected IDIAP to play a major bridging role between AMI and PASCAL.

EURON (European Robotics Network) is another Network of Excellence in the "Future and Emerging Technologies – Beyond Robotics" Strategic Objective, building upon the expertise IDIAP now has in the Brain-Machine Interfaces area.

Other new projects

In late 2002, the US Defense Advanced Research Projects Agency (DARPA) granted IDIAP a small research project in robust speech recognition. As of July 1, 2003, and acknowledging the quality of the collaboration, DARPA significantly increased this funding from approx. 60 K$/year to 270 K$/year to work in the framework of the EARS project (Effective, Affordable, Reusable, Speech-to-text systems, see http://www.darpa.mil/iao/EARS.htm).

Other new projects have been initiated with partners such as Saillabs (Vienna), Infonoia (Geneva), FranceTelecom R&D, etc, and are expected to start in early 2004.

A new agreement with EPFL

On June 2nd, 2003, IDIAP and EPFL signed a new agreement to further strengthen the relationship between the two institutions. The new agreement will further facilitate collaboration at all levels in both academic and research aspects. In particular, the involvement of EPFL in the IM2 NCCR and the academic activities of IDIAP senior staff members now have solid foundations.

Partnership with CIMTEC

The partnership with CIMTEC (www.cimtec.ch), a leading office for business innovation and technology transfer, continues and in 2003 has lead to the creation of a new spin-off company, Spiderphone.ch (see below) and several new projects.

Through this partnership, IDIAP will also play a major role in the new initiative of the Canton du Valais, The Ark, to foster and promote activities in key technological areas such as biotech and IT.

DEWS: Development Economic Western Switzerland

On Thursday May, 1st, 2003 about 30 worldwide correspondents of the DEWS – Development Economic Western Switzerland – visited IDIAP and discovered the past and current research themes of the institute as well as the latest developments in the framework of the IM2 NCCR. This unique opportunity, an immediate consequence of the Canton du Valais recently joining DEWS, means that IDIAP can now rely upon a worldwide network of local contacts. The mission of these contacts is to attract new or existing companies to Western Switzerland (Canton de Neuchatel, Vaud and Valais). To achieve this, they build upon the unique characteristics that our region has to offer, and according to their first impressions, IDIAP and IM2 have great potential to attract companies and research branches in the field.

Spiderphone

In July 2003, a new company Spiderephone.ch was founded as a wholy-owned subsidiary of the US-based Spiderphone Inc. The double objective of Spiderphone.ch is to provide Swiss and European customers the web-enhanced conference call services developped in the US, and to integrate IDIAP technologies in the product offering of both companies. Spiderphone Inc co-founder, Dr Pierre Wellner has been an IDIAP Senior Researcher since 2001 and will guarantee the link between the Institute and the company.

A video to present IDIAP's research areas

A short video illustrating some of the research areas of IDIAP has been produced in collaboration with Sismic, a company based in Monthey. The video is available online on the IDIAP web site, as well as on small CDcards that will play on any computer. The video has both french and english comentary.

Awards

In late 2003, IDIAP had the pleasure to announce two awards that were granted to its researchers: Dr Pierre Weller was awarded the **UIST Lasting Impact Paper Award** for his original DigitalDesk UIST paper published in 1991, and David Grangier received the Eurecom Hitachi award for his internship project at IDIAP.

IDIAP Best PhD Student Award

Since 2002, IDIAP has granted an annual PhD award, based on a selective process taking into account the quality of the scientific research, inter-project collaboration, and the candidate's social qualities. The selection for the prize is made through nomination and selections by all seniors and postdocs. The 2003 IDIAP Best PhD Student award recipient was Ronan Collobert.

Bi-annual PhD student progress reports

IDIAP now collects bi-annual PhD student progress reports which serve as an evaluation and dissemination tool.

Eurospeech'03

IDIAP organized Eurospeech'2003, which was held at the International Congress Centre of Geneva in September 2003. Eurospeech is the premiere conference on speech and language technology, attracting more than 1000 scientists every two years. Details at `http://www.eurospeech2003.org`.

ICASSP Special Session on Smart Meeting Rooms

IDIAP was to chair a Special Session on Smart Meeting Rooms at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), in Hong Kong, April 2003. The session consisted of invited papers by leading research groups, who will present current work on various aspects of this emerging domain. Due to SARS, the whole conference was cancelled, however the proceedings were still published.

Berlin

As a followup to last year's August 1st celebration, which saw a replica of the Matterhorn built in the middle of Berlin, the Canton of Valais was invited by the Swiss Embassy in Berlin for a reception with representatives of the economic, political and diplomatic worlds. After a general presentation by State Councilor Jean-René Fournier, Dr Jean-Albert Ferrez presented IDIAP and the IM2 NCCR, and showcased some research results.

IST Milan

IDIAP, through the M4 and AMI EC projects, was present at the «IST 2003: The Opportunities Ahead» Event which took place during SMAU 2003 in Milan early October. SMAU is a very large computer technology trade show attracting several hundred thousand visitors.

World Summit on Information Society

In the framework of the World Summit on Information Society, CERN has organized an exhibition/conference on behalf of the world's scientific community called SIS-Forum@ICT4D (Science and Information Society Forum at the ICT4D), as a time and a place where a few selected research institutions could meet and present their projects and activities. The SIS-Forum@ICT4D exhibition was displayed at Palexpo, Geneva, from the 9th to the 13th of December 2003. It consisted of digital demonstrations and presentations of projects and activities, all focusing on science's leading role in driving the development of the Information Society.

In this framework, IDIAP was invited to present their research in automatic meeting processing and their current developments in meeting browsers, resulting from national projects (such as IM2) and European projects (such as M4 and AMI). The IDIAP booth attracted a lot of attention from multiple visitors, including BBC who spent about 30 minutes interviewing and recording.

Foire du Valais

The Loterie Romande was one of the guests of the 44th Foire du Valais which took place in Martigny on October 3–12, 2003. It opened a section of its booth to the numerous institutions which benefit from its financial support. As such, IDIAP had the opportunity to present itself and its activities.

Important visits at IDIAP

Several important companies, associations, groups or people have visited IDIAP in 2003:

- The annual meeting of the Directors of Economic Affairs of the 26 Swiss cantons, with the head of SECO David Syz.
- The State Counsilors of the Cantons of Valais and Zug
- A delegation of the Swiss Science and Technology Council
- A delegation of the *Conseil des Ecoles Polytechniques Fédérales*, (CEPF, ETHRat)
- The Committee of the Valais Chamber of commerce and industry
- The Délégation Economique du Conseil Général de Martigny

- Several delegations of Swisscom and its subsidiary Bluewin
- VPs of Logitech
- France Telecom R&D
- The Fondation Suisse pour les Téléthèses
- The members of CCSO, Centre Cim de Suisse Occidentale

5 Research Activities

All IDIAP activities are aiming at the advancement of research, and the development of prototypes, in the field of man-machine interaction. IDIAP is particularly concerned with technologies coordinating natural input modes (such as speech, image, pen, touch, hand gestures, head and/or body movements, and even physiological sensors) with multimedia system outputs, such as speech, sounds, and images. Among other applications, IDIAP is also concerned with multimodal technologies to support human interaction, in the context of smart meeting rooms, such as the one available at IDIAP (www.idiap.ch/ moore/meeting). In that specific context, IDIAP aims to develop new audio, visual and multimodal tools for understanding, searching and browsing meetings data captured from a wide range of devices, as part of an integrated multimodal group communication. IDIAP activities thus address a range of multidisciplinary research including natural speech recognition, speaker tracking and segmentation, visual shape tracking, gesture recognition, multimodal dialogue modelling, meeting dynamics, summarization, browsing and retrieval, and HCI (browser design and browser evaluation). As part of Multimodal Interaction, we also started new research activities in brain-machine interfaces, as briefly discussed in Section 5.6.

With a research staff of more than 60 scientists (including seniors, postdoctoral researchers and PhD students), IDIAP conducts research and developments in several research areas:

1. **Machine learning**: algorithms for classification, regression and density estimation; statistical modelling and classification, connectionist techniques. *Current research activities* include: support vector machines, large-scale optimization problems, kernel methods, expert fusion, ensemble models, theoretical analysis, multi-channel processing, multi-stream HMM, and asynchronous HMM (e.g., merging of different data streams, possibly non-synchronous and with different data rate), multi-modal fusion.

2. **Speech and audio processing**: speech signal processing, multilingual robust speech recognition, development of better speech models, acoustic scene analysis, source localization, microphone array, and low-bit rate speech transmission. *Current research activities* include: improved robustness, better modelling of the time/frequency structure of the speech signal, portability across new applications, language modelling, automatic adaptation (of acoustic and language models), confidence measures, out-of-vocabulary words, spontaneous speech, prosody, modelling temporal dynamics, speaker turn detection (using acoustic features and/or source localization features), microphone array post-processing, low bit rate speech transmission (based on phonetic vocoding).

3. **Computer vision**: object detection, recognition, and tracking, motion analysis, text recognition, gesture recognition, facial expressions, and extraction of relevant information from images and video sequences. *Current research activities* include: object modelling, algorithm robustness, data fusion (color, shape, motion) and feature selection, online learning and model adaptation, multi-object tracking (dynamics and data-likelihood modelling), behavioral models, joint tracking and event recognition, computational complexity.

4. **Information retrieval**: extraction of useful information from multimedia data, audio and video structuring, noisy text retrieval and categorization. *Current research activities* include: content-based information management using multiple data (audio and video) streams, (semi) automated video data description, optimization of user interaction.

5. **Biometric authentication**: speaker verification, face recognition, multimodal user authentication. *Current research activities* include: increasing robustness of user authentication techniques, non-frontal face verification, multimodal fusion, multimodal user authentication (mixture of experts, confidence-based weighting of the different media, etc).

6. **Multimodal Interaction**: HCI design, augmented meeting systems, meeting browser, browser evaluation, visualization, mobile telephone interaction, gesture recognition, brain-machine interfaces, adaptive interfaces. *Current research activities* include: user-composed visualization of recognized meeting activities, design of brain-actuated devices, two handed interaction combined with speech, and phone-based biometric authentication with face detection and tracking.

In 2003, all IDIAP activities were however still organized along the three research groups briefly described below, namely: speech and audio processing, computer vision, and machine learning. Since a lot of this work also took place in the context of the IDIAP Smart Meeting Room (as described in Section 5.4, we also describe some of the specific activities resulting of this research in Section 5.5.

5.1 Speech Processing Group

The overall goals of the IDIAP speech processing group are to research and develop robust recognition and understanding techniques for realistic speaking styles and acoustic conditions, as well as robust speaker verification and identification techniques. 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 projects for several years and is today at the leading edge of technology. The IDIAP Speech Processing group is also involved in numerous national and European collaborative projects, as well as industrial projects.

The IDIAP Speech Processing group is currently involved in numerous European, Swiss National Science Foundation, and DARPA projects, for example:

- MultiModal Meeting Manager (M4), from the EC/IST 5th Framework Program, `http://www.dcs.shef.ac.uk/spandh/projects/m4/index.html`. This project is concerned with the construction of a demonstrable system to enable structuring, browsing and querying of an archive of automatically analysed meetings. The archive will have been created in the Smart Meeting Room installed at IDIAP, equipped with multimodal sensors. See the annual report at `http://www.dcs.shef.ac.uk/spandh/projects/m4/M4-AnnualReport2002/`.
- Hearing, Organization and Recognition of Speech in Europe (HOARSE, `http://www.hoarsenet.org`)
- DARPA EARS (Effective Affordable Reusable Speech-to-text), see `http://www.darpa.mil/iao/EARS.htm`.
- Speaker source localization, microphone arrays and beamforming.

The IDIAP Speech Group is also significantly contributing to the Swiss National Center of Competence in Research (NCCR) IM2 (see `http://www.im2.ch`) on Interactive Multimodal Information Management, through "Individual Projects" IM2.SP (`www.im2.ch/SP.php`) and IM2.ACP (`www.im2.ch/ACP.php`).

5.1.1 Research Themes

The research areas of the Speech Processing group currently focus on:

- Automatic recognition of (isolated, continuous, or natural) speech based on phonetic (sub-word) modelling, using spectral-temporal profiles of speech, as well as articulatory.
- Development and improvement of state-of-the-art speech recognition systems based on hidden Markov models (HMM).
- Speaker verification: development and improvement of state-of-the-art text-dependent, text-independent, and user-customized speaker verification systems.
- Using discriminant artificial neural networks (ANN) to estimate a posteriori probabilities. In this regard, IDIAP (in collaboration with ICSI, Berkeley, `http://www.icsi.berkeley.edu`) is recognized as a leader in the use of hybrid HMM/ANN systems, exhibiting several advantages compared to standard HMM approaches.
- Estimation of confidence levels, i.e., attaching a confidence score to each recognized word to indicate how likely the word is correctly recognized. In this context, the problem of detecting out-of-vocabulary words is also investigated.

- Multi-stream and multi-band speech recognition: improving robustness of state-of-the-art systems based on multiple feature streams. This includes the extraction of multiple features from the same input utterance, exhibiting different properties, such as multiple temporal resolutions and/or containing some new, novel, or robust type of information. As a particular case, multi-band speech recognition, combining multiple (HMM or HMM/ANN) recognizers, has been shown to significantly improve robustness to narrow band noise.

- Multi-stream and multi-channel combination: Developing novel methods to combine information generated from multiple experts trained on multi-stream features to improve word recognition and increase robustness of the recognition to corrupting environmental conditions.

- Acoustic change detection and clustering, as required when dealing with large audio and multimedia databases (such as broadcast news and sport videos). In this framework, different approaches are investigated towards automatic segmentation of (multimedia) sound tracks, including, among others, changes in acoustic environments, speaker change detection, speaker identification and tracking, and speech/music discrimination. This segmentation is also useful, e.g., towards automatic adaptation of the models, as well as for resetting time points for language models and topic extraction systems.

- Pronunciation variants modelling: Automatic extraction and modelling of pronunciation variants based on various factors such as word context and speaking style (e.g., conversational speech, speaking rate).

- Statistical language modelling: Extending current language models to better cope with natural speech, out-of-vocabulary word, and word classes.

- Speaker adaptation: Improving recognition accuracy by automatically adapting (a subset of) the parameters of the recognition system.

- Development and adaptation of efficient software for large vocabulary continuous speech recognition, on different computer platforms (mainly UNIX and Windows NT), all compatible with the TORCH (`http://www.torch.ch`) libraries developed at IDIAP.

- Speaker source localization, microphone arrays and beamforming, as illustrated in Figure 1.

- Development and testing of applications prototypes.

5.1.2 Application Examples

1. Command and control systems, possibly used in noisy environments, e.g., to operate a speech enabled cellular phone in cars. See, e.g., the RESPITE and SPHEAR projects.

2. Speech enabled information systems: Building speech-enabled kiosks, desk tablets, and personal data assistants to enable users to find and display current information.

3. Information retrieval for audio documents: Using transcriptions automatically generated by a large-vocabulary speech recogniser to build indexes that can be queried by information retrieval engines for searchable audio archives. See, e.g., the ASSAVID (see Figure 2) and CIMWOS projects, allowing for:

    - Automatic transcription of broadcast speech by an automatic speech recognition system
    - Automatic indexing of the generated audio archives
    - Content-based retrieval from typed or spoken input queries.

4. Automatic meeting manager: processing of multiple audio (and video) streams for structuring, browsing and querying of an archive of automatically analyzed meetings. See, e.g., the M4 project or a typical meeting browser.

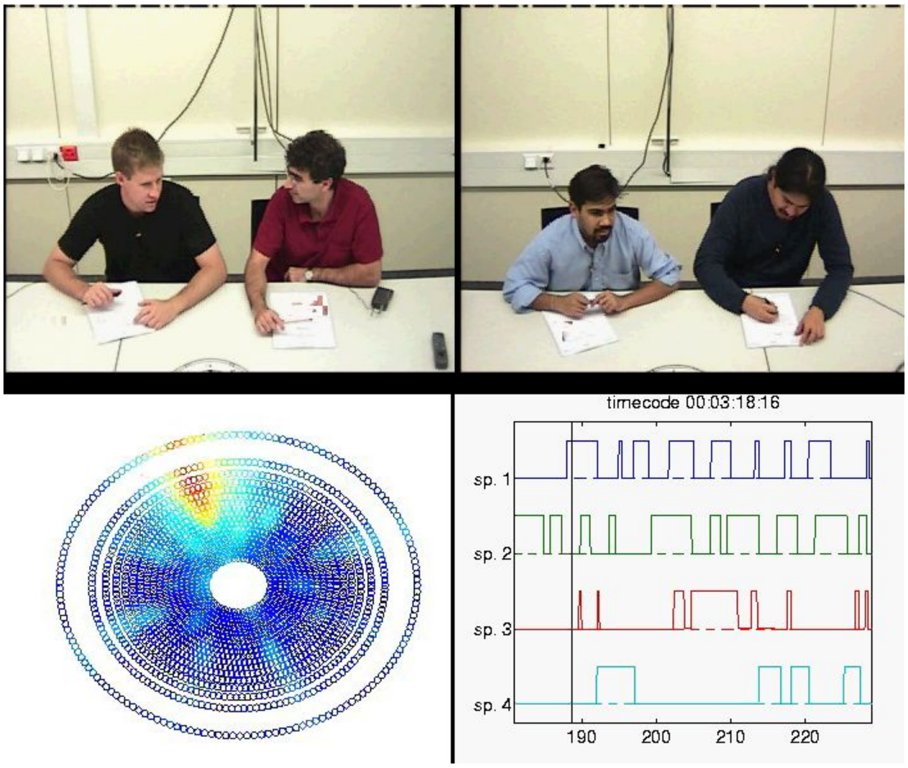

Figure 1: Illustration of the audio tracking and speaker beamforming achieved by using the microphone arrays installed in the IDIAP smart meeting room.

5.2 Computer Vision Group

The computer vision group at IDIAP investigates and develops principled methods and algorithms for analysis of visual and multimedia data, and addresses a number of specific problems including people detection and tracking, gesture and facial expression recognition, handwriting recognition, and multimedia content analysis. Their work frequently involves collaboration with the two other groups at IDIAP, speech processing and machine learning, as complementary expertise is needed for many of the research problems. The group is active in all of their areas of expertise under a number of collaborative European and Swiss national projects.

5.2.1 Research Themes

1. Handwriting recognition: Offline handwriting recognition is the automatic transcription of handwritten data when only its image is available. Members of the group have worked on methods to improve modelling of handwritten data, and have developed a recognizer based on continuous density HMMs which can deal with single words as well as handwritten texts (with the help of Statistical Language Models).

2. Face Algorithms: Face algorithms can be divided into four different areas.

   - Face detection: The goal of face detection is to identify and locate human faces in images at different positions, scales, orientations and lighting conditions.
   - Face localization: Face localization is a simplified face detection problem with the assumption that the image contains only one face.
   - Face verification: Face verification is concerned with validating a claimed identity based on the image of its face, and either accepting or rejecting the identity claim.

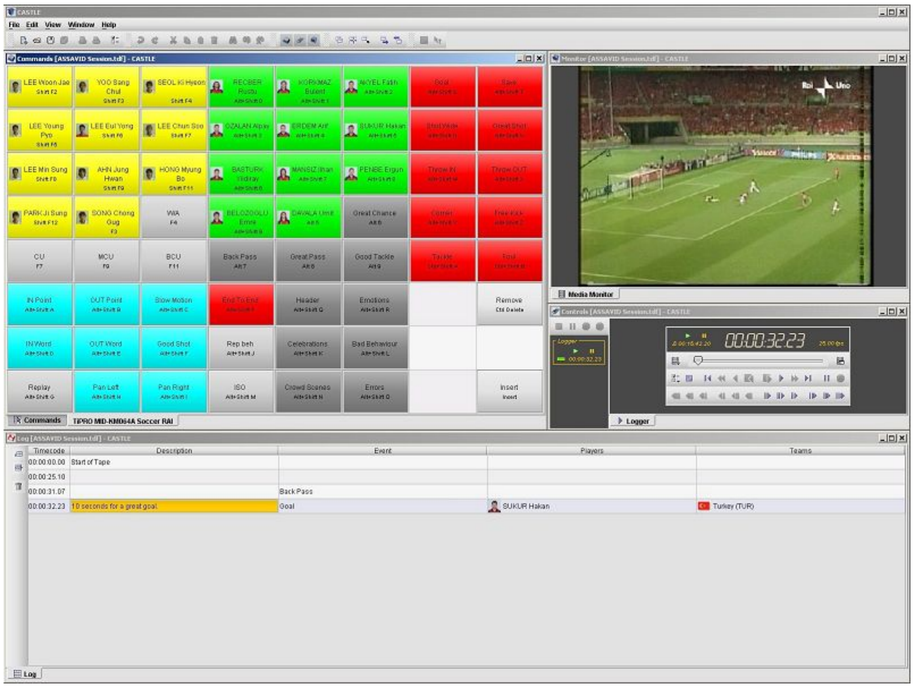

Figure 2: Interface of the audio-video indexing and retrieval system developed in the framework of ASSAVID.

- Face recognition: The goal of face recognition is to identify a person based on the image of its face. This face image has to be compared with all registered persons. Therefore, face recognition is computationally expensive with respect to the number of registered persons.

The group is mainly interested in face detection and verification using neural networks, SVM based methods and boosted weak classifiers (see Figure 3).

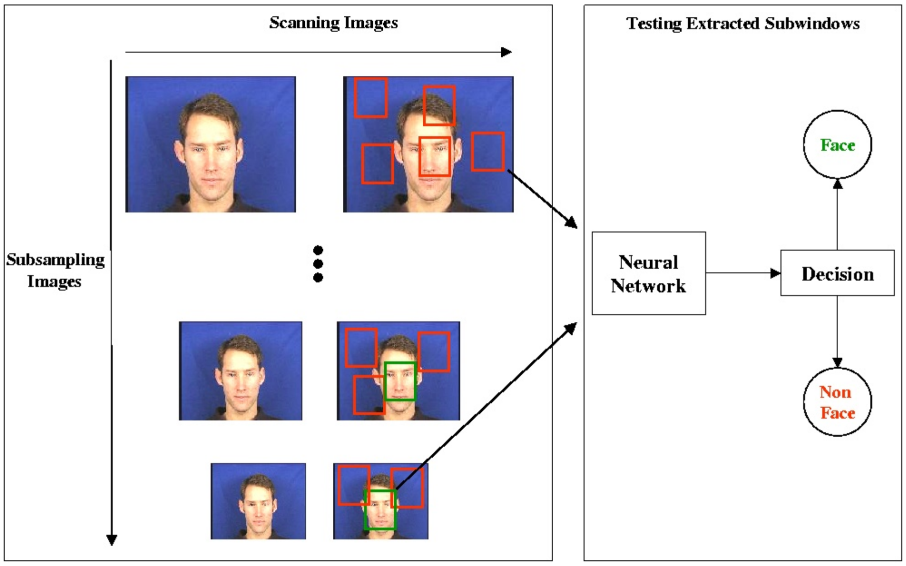

Figure 3: Face detection in an image using several subsampling stages.

3. Gesture Recognition: Gestural interaction based on the image is the most natural method for the construction of advanced man-machine interfaces. Thus, machines would be easier to use by associating the gestural command with the vocal command. This includes recognition of gestures such as facial expressions, hand postures, hand gestures and body postures. Current work on facial expression recognition is based on convolutional neural networks. Statistical approaches (skin color blobs) for object segmentation (faces and hands) in color images are investigated. The vision group is also interested in gesture recognition using hybrid models (Hidden Markov Models and Neural Networks) such as Input/Output Hidden Markov Models.

4. Tracking and activity recognition: Object tracking represents an essential component of gesture recognition, human behavior monitoring, and video indexing. IDIAP is investigating the design of stable trackers that are robust against ambiguities, image measurements, changes in the acquisition setting, and object intra-class variability. The group focuses on two areas: (1) the development of sequential Monte Carlo (SMC) techniques, and the combination of SMC and finite state motion models based on HMMs for joint tracking and recognition of people activity, and (2) the fusion of multiple visual and multimodal (audio-visual) features, for example for speaker tracking.

5. Multimedia content analysis: The vision group is developing statistical models, algorithms and tools to automatically extract relevant information from audio-visual streams, which can be used for structuring, annotating, indexing and retrieving multimedia databases. Some of the current research directions include:

   - Media structuring: The structure of videos is needed both at the individual and at the database levels. On one hand, finding structure in individual videos (shots, scenes) is useful to generate video summaries for browsing and retrieval, and usually constitutes the starting point to extract higher-level information. On the other hand, structuring a whole video database is useful for access and filtering (locating video replicas, organizing by "visual topic", etc.).
   - Event classification: The group is developing audio-visual feature extraction and data fusion algorithms for event classification in sports video and meeting databases. Current efforts have been directed to define semantically meaningful events, and to learn their statistical models for classification.
   - Text Detection and Recognition in Images and Videos: The vision group is involved in text detection and segmentation algorithms, and also examination of new paradigms in video text recognition. The goal of current research is to exploit the temporal redundancy to fuse recognition results of the same text obtained at different times.
   - Modeling of textual and visual features: Members of the group investigate joint statistical models of words and visual features in multimedia databases, to relate low-level visual information with semantics. Such models would allow for important information retrieval functionalities, like clustering (grouping images that refer to the same text topics), annotation (attaching words to visual content), and illustration (attaching images to words).

5.2.2 Application Examples

Three typical applications of the methods developed at IDIAP are the following:

- Hand drawn character and cursive writing recognition is useful for such tasks as automated address reading for postal services, and interfaces for such devices as PDAs. In addition, notes taken in meetings and during other discussions are predominantly handwritten. The ability to read such sources of information would be highly useful in many cases.

- Identity verification is a general task for security applications like access control, transaction authentication (in telephone banking or remote credit card purchases), voice mail, or secure teleworking. Face detection is a fundamental step before the verification procedure. Its reliability and time-response have a major influence on the performance and usability of the whole face verification system.

- The purpose of image and video annotation is to provide access to the ever increasing digital archives of such data. Whether these archives are within a television station or publicly available web documents, the volume of data being produced at any moment is beyond human ability to annotate. In addition there are large historical archives that contain priceless data recording important moments. Television stations will use such technology to provide a method of access to their archives, such as sports and news, and to access historical footage to enrich current programs, and for documentary pieces. Video and image text recognition is obviously a key part of this technology, as captions and in-vision text contain much useful information.

5.3 Machine Learning Group

The Machine Learning group at IDIAP is mainly interested in statistical machine learning, a research domain mostly related to statistical inference, artificial intelligence, and optimization. Its aim is to construct systems able to learn to solve tasks given a set of examples that were drawn from an unknown probability distribution, eventually given some prior knowledge of the task. Another important goal of statistical machine learning is to measure the expected performance of these systems on new examples drawn from the same probability distribution.

5.3.1 Research Themes

1. Large scale data analysis: most actual powerful machine learning algorithms have been used for medium scale datasets: less than one hundred features describing one example and less than ten thousand example in the dataset. For instance, the now well-known Support Vector Machine algorithm needs resources that are quadratic in the number of examples, which forbid their use for problems with more than a few hundred thousands examples. Decomposition of the problem into sub-problems may lead to efficient solutions.

2. Ensemble models: One way to enhance generalization performance of machine learning algorithms is to combine the output of many algorithms instead of relying on only one algorithm. Many such methods are already known, such as AdaBoost, Bagging, Mixture of Experts.

3. Feature selection: Another way to enhance generalization performance of machine learning algorithms is to select and use only the input features that are well suited to solve a given problem.

4. Fusion of generative and discriminative models: two classes of machine learning algorithms are known and they have different advantages and disadvantages, depending on the problem to solve. We are interested in new algorithms that take advantages of both approaches.

5. Generalization performance analysis: As already stated, the goal of our group is not only to provide new and efficient machine learning algorithms but also to analyze and understand them in order to be able to compare them to other state-of-the-art algorithms.

6. Sequence modelling: most recent machine learning algorithms have been tailored for static problems. Given IDIAP's interest in speech processing, our group is also interested in developing and analyzing specific machine learning algorithms for sequence processing, including time series prediction and biological sequence analysis.

7. Spatial data analysis: We are specifically interested in building machine learning algorithms that would take into account spatial correlation between the input features and the target output in order to simultaneously enhance the prediction performance while preserving the spatial distribution of the dataset.

8. Multi-class classification: Many machine learning algorithms are in fact classification problems with multiple classes. One such problem in speech is the prediction of the phoneme (one out of 40 different phonemes) given the input features, at every time step.

9. Brain-Machine Interfaces (BMI): Using specially design helmets, EEG signals of a patient can be recorded and analyzed by advanced machine learning techniques, in order to extract corresponding commands uttered mentally by the patient (such as "left", "right", etc). Both high-level and low-level processing of these very noisy sequences are taken into account, from simple FFTs to Hidden Markov Models.

10. Support to the Vision and Speech groups: the main role of the machine learning group is to support the research of the two other groups when machine learning is concerned.

5.3.2 Application Examples

The applications of Statistical Machine Learning are quite diverse. On top of all the applications related to speech and vision, which are best described by the two other groups, here is a sample of other interesting application domains:

- Data Mining: how to extract interesting information from huge database wharehouses (for instance, churn detection, client modeling and prediction).

- Finance and Economy: financial portfolio management, asset prediction, portfolio selection, auction analysis.

- Pattern Recognition: handwritten character recognition, speech recognition, face detection.

- Biological Sequence Analysis: classification of DNA or RNA sequences.

5.4 IDIAP Smart Meeting Room

Figure 4: Set up of the IDIAP Smart Meeting Room allowing for synchronized capture of audio (24 channels), video (3 channels), white board activity, PC-projector, and handwritten notes

Figure 5: Capture of handwritten notes by using Logitech pens.

5.4.1 Overview

In the scope of IM2 (as well as a new, related European project, referred to as MultiModal Meeting Manager "M4", see http://www.dcs.shef.ac.uk/spandh/projects/m4/) IDIAP has equipped a meeting room with audio and video acquisition facilities for recording and processing meetings. See http://www.idiap.ch/ moore/meeting/ for full details.

The smart meeting room hardware is designed to be capable of:

- Recording the speech signal of each meeting participant using both close-talking microphones and table-top microphone arrays.
- Recording multiple video cameras, including wide angle and medium-range shots of the participants.

To facilitate fusion of modalities in some processing tasks, the audio and video systems are synchronized using a master sync signal, and each recorded channel is accurately timestamped.

5.4.2 Audio Acquisition Hardware

The audio hardware system consists of 24 microphone channels that are digitised and streamed directly to the hard disk of a PC. The specific hardware & software components are :

- 24 high quality Sennheiser MKE2-5-C lapel microphones
- Custom-built microphone power box
- 1 Sennheiser 100 series wireless microphone transmitter/receiver

- 3 PreSonus Digimax preamplifiers/digitisers (8 channels each)
- 1 MOTU 2408 mkII PC audio interface
- PC (Pentium 4-2GHz with 2 x 80 Gb hard disk, DVD Writer, Windows 2000 Pro)
- Cakewalk SONAR recording software

All audio channels are digitised at 48kHz with 24-bit resolution.

5.4.3 Video Acquisition Hardware

The video acquisition system consists of 3 video channels that are each recorded by separate MiniDV video tape recorders. This ensures that all acquired video is high-quality digital and that full PAL framerate and resolution is preserved.

The video acquisition hardware components are :

- 3 Sony SSC-DC58AP cameras
- 3 wide angle lenses
- 3 Sony GV-D1000E MiniDV video walkman

5.4.4 Synchronisation

In order to playback video & audio recordings and to facilitate multimodal processing tasks, it is crucial that all video and audio acquisition is synchronised and that the recordings are accurately timestamped.

All cameras are frame-locked using a master blackburst sync signal, and a timecode that is also synchronised with the master sync is generated and added to the audio and video recordings. Other timing signals, such as the 48kHz clock used for audio digitisation, are also derived from the master sync signal.

The synchronisation hardware components are :

- 1 Horita BSG-50 blackburst generator
- 1 MOTU MIDI Timepiece AV timing control module
- 3 Horita AVG-50 Active VITC inserters

5.4.5 Installation

The physical properties of the IDIAP Smart Meeting Room are :

- Rectangular table with seating for 12 people
- Whiteboard, projector screen, [and beamer, still to be installed]
- Carpet floor and uneven ceiling for reduced reverberation
- Fluorescent lighting
- Air conditioning (not installed yet)
- Ceiling rails for mounting cameras and lights

The audio acquisition hardware, synchronisation hardware and MiniDV recorders are housed in a 19" rack that sits at the rear of the room. The underside of the meeting room table has channelling, which hides all wiring and provides each seating position with a socket where each participant can plug his/her lapel microphone.

Four 30cm linear microphone array tabletop mounts have been constructed. These are designed to mount 4 or 8 microphones and are placed in the center of the table between 2 oppositely seated meeting participants.

2 fully adjustable camera mounts have also been constructed, and are suspended from the ceiling rails. The third camera has been temporarily mounted on a tripod, and will be suspended from the ceiling when an optimal camera configuration has been determined.

5.4.6 Recording Meetings

For full audio-video meeting recordings, the Smart Meeting Room can accommodate 6 meeting particpants in its current configuration. Two cameras each provide a medium-angle front-on view of 3 participants, while the 3rd tripod-mount camera sits on the table and provides a wider angle view of the entire meeting, including the whiteboard and projector screen. Each meeting participant wears a lapel microphone, and 4 to 8 microphones are used in each of 3 tabletop microphone arrays.

For audio-only recordings, the Smart Meeting Room can accommodate up to 12 meeting participants.

5.5 Research Activities Related to Smart Meeting Room

5.5.1 Audio-visual tracking

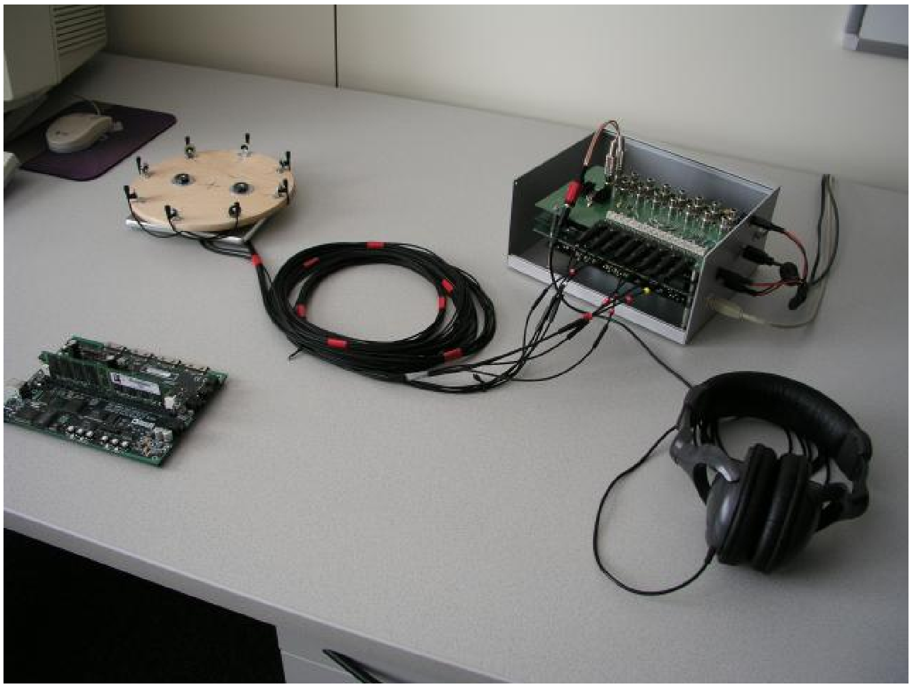

Figure 6: Small microphone arrays, and its realtime hardware, typically used in the IDIAP smart meeting room to perform acoustic source localization.

We have developed a principled method for speaker tracking, fusing information coming from multiple microphones and uncalibrated cameras, based on *Sequential Monte Carlo* (SMC) methods, also known as *particle filters* (PFs). For a state-space model, a PF recursively approximates the conditional distribution of states given observations using a dynamical model and random sampling by (i) generating candidate configurations from the dynamics (*prediction*), and (ii) measuring their likelihood (*updating*), in a process that amounts to random search in a configuration space. Data fusion can be introduced in both stages of the PF algorithm.

Our work is guided by inherent features of AV data. First, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns across cameras), and (re)initialisation. Its use for prediction thus brings benefits to modelling real situations. Second, audio can be inaccurate at times, but provides a good initial localization guess that can be enhanced by visual information. Third, although audio might be imprecise, and visual calibration can be erroneous due to distortion in wide-angle cameras, the joint occurrence of AV information in the constrained physical space in meetings tends to be more consistent, and can be learned from data. See project IM2.MUCATAR for more detail.

5.5.2 Media File Server

In 2003, IDIAP developed the Multimodal Media File Server (see http://mmm.idiap.ch), which is a repository for storing audio and video recordings to support research on multimodal information processing. The initial data available is a set of meeting recordings taken from the IDIAP smart meeting room. Each recording consists of output from three cameras at full PAL resolution and frame rate, plus audio from at least one microphone array and lapel microphones for each participant.

As illustrated in Figure 7, the server provides HTTP, RTSP, and FTP interfaces to support browsing, playing, retrieving, and adding of recorded multimodal data files. It can also serve as a platform to support future browsing and searching applications. This Media File server provides the following functionalities:

- Browsing: Each recorded session has a directory on the file server and a dynamically generated "home page" that displays all available files including a jpeg image for each video file. Every file is downloadable from this page by FTP or HTTP.

- Playing: The session "home page" has images and buttons to stream any audio or video file using RealPlayer on Unix or Windows, and a "Synchronized Play" button that dynamically generates a SMIL presentation from all user-selected "checked" media clips to display them simultaneously in sync. Start time offsets and durations are part of the file names to allow the media file server (and other software) to account for varying start times of concurrently recorded audio and video files.

- Retrieving: Data can be downloaded for processing on your local computer using either HTTP or FTP.

- Adding: Tools are available to rename and format recorded data so it can be displayed, previewed, and retrieved using the web and SMIL user interfaces on this server. At this time, however, new data files are still added manually by server administrators.

The Media File Sever provides capabilities for browsing, playing, and retrieving recorded meeting data on line, and it has been in use by researchers (from IM2 and worldwide) since mid January 2003.

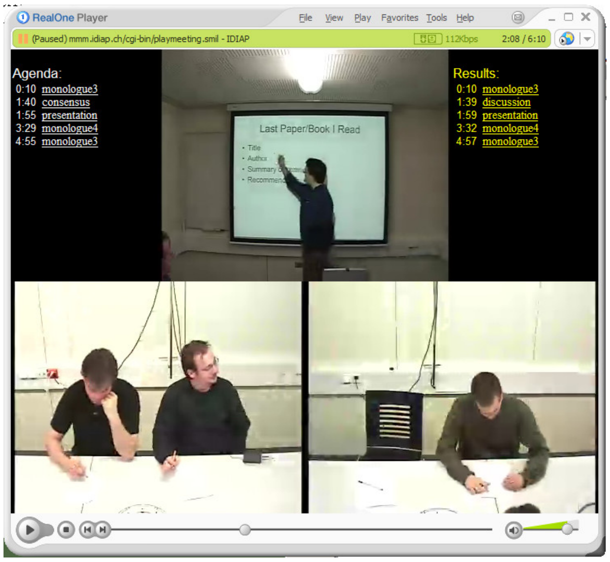

Figure 7: Synchronized output presentation, including audio, video and text annotation. In the case presented here, the hand annotation is presented on the left side of the central view, while the automatic annotation is given on the right side.

5.5.3 SMR Data Recording

In the 12 months since the IDIAP multimodal meeting room became operational, an initial corpus of multimodal meetings has been recorded. Meetings were recorded using 3 cameras and 12 microphones, with all channels fully synchronized. Currently, the database contains 60 meetings, where each meeting consists of 4 participants and lasts approximately 5 minutes. The meetings are loosely scripted in terms of the type and schedule of the high-level actions, but otherwise the content is natural. The corpus is fully described in several papers and is currently being expanded and made available for (worldwide) public distribution through the Multimodal Media File server (see below). The audio streams of this data have also been transcribed at the word level (using the ChannelTrans tool) and is also publicly available. Several IM2 partners, as well as numerous international laboratories, have already started working on this data.

IDIAP is now preparing to record a new meeting corpus that addresses some of the limitations of our existing corpus. Limitations of this initial corpus were discussed within IM2, as well as with other European partners, and a draft specification for new meeting data collection has been prepared and is currently being shared for comments with all the partners.

5.5.4 Processing of Multimodal Data

Figure 8: People tracking in the IDIAP smart meeting room.

Thanks to the above developments, IDIAP was able to make significant progress in key areas of multimodal processing, such as:

1. *Multimodal group actions in meetings*: automatic segmentation and identification of multimodal group actions, which already resulted in several publications, and which is attracting a lot of interest worldwide.

2. *Audio-visual speaker tracking*: where (particule filter based) visual (people) trackers could be improved and initialized by using the information contained in the associated audio channels, resulting in a truly multimodal, and adaptive, audio-visual tracker. We developed a method that fuses information coming from multiple microphones and uncalibrated cameras based on particle filtering algorithms. Our work was guided by several inherent features of AV data. First, audio is a strong cue to model dynamic discontinuities, like speaker turns across cameras and re-initializations. Second, although audio can be inaccurate, it usually provides good localization candidates that can be refined with visual information. Third, the joint occurrence of AV information can be used to calibrate the sensors using relatively simple procedures, despite the inaccuracy in individual modalities. Our approach exploits these features in a principled way, via mixed-state i-particle filters. The methodology has involved the development of (1) an audio localization algorithm that detects speaker changes with low latency, while maintaining good estimation accuracy, and (2) a probabilistic model that integrates models of people dynamics, shape, color, and sound observations, and a mechanism for camera switching. The methodology was tested in the SMR, and showed the ability to initialize and track moving speakers, and switch between speakers across cameras, while tolerating visual clutter. Extensions of the methodology towards multi-speaker tracking, with enhanced observation models, were also initiated. Finally, a benchmark dataset for audio localization, part of which will be used to test AV tracking algorithms, was recorded in the SMR, and is currently being annotated.

5.5.5 Meeting Browser

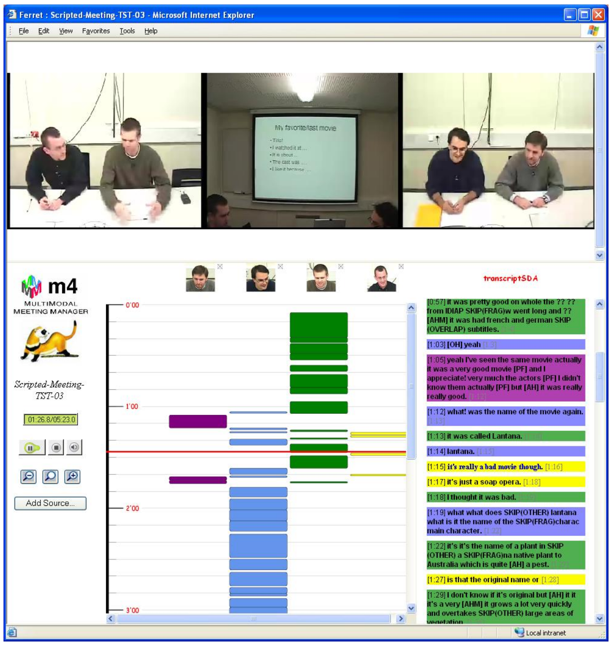

Figure 9: Typical scenario and interface for a meeting browser.

An experimental browser, called Ferret (see Figure 9), allows users to select any combination of available data streams as XML files, and to display them alongside each other for inspection and comparison. Graphical representations of interval streams are displayed on a scrollable and zoomable timeline in the lower part of the browser, while the upper part is for synchronized playback of audio and video. Clicking on elements in the timeline below controls media playback and a red horizontal cursor line that moves down along the timeline. Additional interval data streams (stored as XML files on any web server) can be added or removed from the timeline pane at any time, and text transcripts of the meeting can also be displayed when available. This browser will be used to help us evaluate which kinds of automatic segmentation data are most useful in helping people to find points of interests within a recorded meeting.

5.6 Brain-Machine Interfaces (BMI)

Figure 10: IDIAP brain-machine interface.

During 2003, we have set up a first version of the brain-machine interface and different sets of EEG data have been recorded from five subjects using several protocols. Most of the research effort has focused on the recognition of 3 mental tasks the subject concentrates on asynchronously; i.e., she can switch from a task to another at any time. This allows faster and more natural operation of brain-actuated devices. The mental tasks we have analyzed are imagination of self-paced movements of the left and right hands, and mental generation of words starting with a random letter. For classification, we have explored the use of Input-Output Hidden Markov Models (IOHMM) to exploit the dynamics of brain activity and also on-line learning algorithms to adapt on the fly the classifier to the its individual user. Results show that IOHMM performs better than other Markovian models, while on-line adaptation improves significantly classification rates between sessions. In addition, we have conducted a deeper analysis of the results achieved by the subjects who learned to control our previous brain-actuated devices, a mobile robot in particular.

6 Current Projects

AMI – Augmented Multi-party Interaction

Funding: European Integrated Project, 6th Framework Programme, Information Society Technology, supported by OFES

Duration: January 2004 – December 2007

Partners: IDIAP (Coordinator), DFKI (D), ICSI (Berkeley, USA), TNO (NL), Brno University (CZ), Technical Univ. of Munich (D), Univ. of Edinburgh (UK), Univ. of Sheffield (UK), Univ. of Twente (NL), FastCom (CH), Novauris (UK), Philips (NL), RealVNC (UK), Spiderphone (CH), W3C (F)

Contact persons: Hervé Bourlard, Iain McCowan

Description: AMI (`www.amiproject.org`)is concerned with new multimodal technologies to support human interaction, in the context of smart meeting rooms and remote meeting assistants. The project aims to enhance the value of multimodal meeting recordings and to make human interaction more effective in real time. These goals are being achieved by developing new tools for computer supported cooperative work and by designing new ways to search and browse meetings as part of an integrated multimodal group communication, captured from a wide range of devices.

Integral, multi-disciplinary, research areas: The present Integrated Project (IP) is thus very ambitious and addresses a wide range of critical multi-disciplinary activities and applications, covering:

1. Multimodal input interface: including multilingual speech signal processing (natural speech recognition, speaker tracking and segmentation) and visual input (e.g., shape tracking, gesture recognition, and handwriting recognition).
2. Integration of modalities and coordination among modalities, including (asynchronous) multichannel processing (e.g., audio-visual tracking) and multimodal dialogue modelling.
3. Meeting dynamics and human-human interaction modelling, including the definition of meeting scenarios, analysing human interaction and multimodal dialogue modelling.
4. Content abstraction, including multimodal information indexing, summarising, and retrieval.
5. Technology transfer through exploration and evaluation of advanced end-user applications, evaluating the advantages and drawbacks of the above functionalities in different prototype systems.
6. Training activities, including an international exchange programme.

AudioSkim – Automatic Segmentation of Large Audio and Multimedia Documents

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Jitendra Ajmera, Iain McCowan, Hervé Bourlard

Description: The problem of distinguishing speech signals from other audio signals (e.g., music) has become increasingly important as automatic speech recognition (ASR) systems are applied to more and more real-world multimedia domains. Furthermore, audio and speech segmentation will always be needed to break and structure the continuous audio stream into manageable chunks applicable to the configuration of the ASR system.

This project thus aims at developing and testing on large audio databases (possibly as part of multimedia databases) such as broadcast news and sport videos, different approaches towards automatic segmentation of (multimedia) sound tracks, including, among others, changes in acoustic environments, speaker change detection, speaker identification and tracking, and speech/music discrimination. During its first year, this project resulted in an automatic system allowing for online segmentation of an audio signal into speech/non-speech segments and which is apparently outperforming other state-of-the-art approaches (see previous activity report). In 2002, the main emphasis of AudioSkim has been put on the automatic (unsupervised) speaker clustering and speaker turn detection. In this framework, a threshold free (BIC like) algorithm was tested and evaluated on big database (Hub-4 97 evaluation set). The results obtained are comparable to the best results when BIC is used with optimal (manually optimized, database dependent) threshold/penalty term. This will be a key development in many applications, including: meeting data segmentation and indexing, multimedia database segmentation and indexing, etc. This has led to several publication; see, e.g., IDIAP Research Reports 02-39 and 02-23.

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

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

Duration: February 2000 – May 2003

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

Contact persons: Samy Bengio, Sebastien Marcel, Johnny Mariethoz

Description: The objectives of the project are to develop and 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.

BN-ASR – Modeling the hidden dynamic structure of speech production in a unified framework for robust automatic speech recognition

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Todd Stephenson, Andrew Morris, Hervé Bourlard

Description: The main objective of this project is to develop new acoustic/phonetic models of speech for Automatic Speech Recognition (ASR). For years, Hidden Markov Models (HMM) have been the most successful technique in ASR. However, HMMs are rather general purpose stochastic models that only crudely reflect the nature of speech. This project will extend the hidden 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 the past and current developments of the Bayesian networks theory. It is expected to contribute to this area as well.

This 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.

CIMWOS – Combined IMages and WOrd Spotting

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

Duration: April 2001 - October 2003

Partners: Institute for Language and Speech Processing (ILSP, Greece), KULeuven (BE), ETHZ (CH), Sail-Labs (Austria), Canal+ (BE), and IDIAP

Contact persons: Iain McCowan, Jean-Marc Odobez, Jitendra Ajmera, Hervé Bourlard

Description: This project aims to facilitate common procedures of archiving and retrieval of audio-visual material. The objective of the project is to develop and integrate a robust unrestricted keyword spotting algorithm and an efficient image spotting algorithm specially designed for digital audio-visual content, leading to the implementation and demonstration of a practical system for efficient retrieval in multimedia databases. Specifically, a system will be developed to automatically retrieve images, video, and speech frames from an audio-visual database based on keywords entered by the user through keyboard or speech. Combined word and image spotting will be used and will provide an efficient mechanism enabling focused and precise searches with improved functionality and robustness. The CIMWOS system aims to become a valuable assistant in promoting the re-use of existing resources thus cutting down the budgets of new productions.

COST 275 – Biometrics-Based Recognition of People over the Internet

Funding: European project, 5th Framework Programme, COST, supported by OFES

Duration: June 2001 - May 2005

Countries involved: Belgium, Denmark, France, Ireland, Italy, Portugal, Spain, Slovenia, Sweden, Switzerland, Turkey, United Kingdom

Contact persons: Samy Bengio, Hervé Bourlard

Description: The main objective of the action is to investigate effective methods for the recognition of people over the Internet based on biometric characteristics (principally voice and facial) in order to facilitate, protect, and promote various financial and other services over this growing telecommunication medium. In operational terms, the main objectives can be specified as follows:

1. To improve knowledge of the issues and problems involved.
2. To study the current techniques for voice and face recognition and to evaluate their performance in the medium considered.
3. To investigate methods for the fusion of the considered biometrics data and the interpretation of the results.

4. To analyze the implementation problems including user-interface issues and investigate effective solutions.
5. 5. To identify the potential applications and analyze the requirements of these.
6. 6. To develop standard methods and tools for the assessment of biometrics-based identification methods.

The secondary objectives are as follows:

1. To promote further research into (a) new and effective methods for voice and face recognition, and (b) novel techniques for data fusion.
2. To further research into multilingual interactive systems and their applications.
3. To standardize methods for the identification of individuals over the Internet.
4. To study the requirements and preferences of industry, and the attitude of the consumers.

As a partner of the COST 275 Action, IDIAP will be active in most of the research themes of the Action, with a particular emphasis on speaker recognition, face recognition, data fusion and assessment. However, thanks to the present project, these activities will take place in the framework of common efforts towards the research and development of a truly multi-modal (using voice and face characteristics) user authentication systems, with applications to internet transactions.

COST 278 – Spoken Language Interaction in Telecommunication

Funding: European project, 5th Framework Programme, COST, supported by OFES

Duration: June 2001 - May 2005

Countries involved: Belgium, Switzerland, Czech Republic, Germany, Spain, Finland, France, Greece, Hungary, Italy, The Netherlands, Norway, Portugal, Sweden, Slovenia, Slovakia, Turkey, United Kingdom

Contact persons: Hervé Bourlard, Sébastien Marcel

Description: The main objective of the proposed action is to "increase the knowledge of potentially useful applications and methodologies in deploying spoken language interaction in telecommunication. Emphasis is on achieving knowledge of speech and dialogue processing in multi-modal communication interfaces". Furthermore, the objective is to achieve knowledge of natural human-computer interaction through more cognitive, intuitive and robust interfaces, whether monolingual, multi-lingual or multi-modal. In operational terms, the main objectives can be specified as follows.

1. To improve the knowledge of the issues and problems involved in general in spoken language interaction in telecommunication.
2. To achieve knowledge of issues related to robustness and multi-linguality within spoken language processing.
3. To achieve knowledge of spoken language interaction in the context of multi-modal communication.
4. To achieve knowledge of human-computer dialogue theories, models and systems and associated tools for the establishment of such systems.
5. To achieve knowledge of and evaluate telecommunication applications that apply spoken language as one out of more input or output modalities.

As a partner of the COST 278 Action, IDIAP will mainly contribute to the Speech Input Processing and Multi-Modal Processing Working Groups. While these two research themes will address several open issues, the present project will also allow us to investigate these issues in the same general framework of robust multi-stream/multi-channel processing, as recently pioneered by IDIAP. The project will also allow further developments of related technologies in robust speech recognition and selected computer vision approaches such as the recognition of pointing gestures and face detection.

EARS - Effective Affordable Reusable Speech-to-text

Funding: DARPA - US

Duration: July 2002 - June 2007

Contact persons: Hervé Bourlard

Description: As part of the DARPA EARS (Effective Affordable Reusable Speech-to-text) program, and in collaboration with ICSI/Berkeley, SRI, the University of Washington, and Columbia University, we are working towards significanlty improving speech recognition in a project refered to as "Pushing the Envelope - Aside" where we are studying both replacements of the standard spectral envelope as the speech representation of choice (typically with cepstral transformation). This includes work on the acoustic "front end", but also includes research on statistical modeling for the new features that are being generated.

Ensembles for Sequence Processing

Funding: Swiss National Science Foundation, now part of the global MULI project

Duration: see MULTI

Contact persons: Christos Dimitrakakis, Samy Bengio

Description: Ensemble methods, such as mixtures of experts, AdaBoost and bagging had originally been developed for classification and regression problems. In both classification and regression, the target is represented by a single, fixed-length, real-valued vector for each presented example. However, in some applications, the target is a sequence of symbols. One such application is speech recognition, where the hypothesis is a sequence of words. The most common machine learning algorithm for sequence processing employs Hidden Markov Models (HMMs) - however, little research has been done in designing new ensemble learning algorithms that are specific to HMMs, or, more generally, to sequence processing.

Research includes studying, proposing, developing and comparing new ensemble methods tailored to sequence processing problems. As the application of ensemble methods has usually resulted in performance improvement in both classification and regression problems, it is expected that it may also increase the efficiency of machine learning on sequence processing problems. One particular area of interest that will be explored is the combination of a set of sequence hypotheses, varying in length and confidence, into the most likely sequence. Another important area of research is related to the sampling of the input space with respect to the training of the base models. The sampling is most problematic in the case where the input consists of a continuous sequence, rather than being segmented into sub-sequences.

EURON – European Robotics Network

Funding: European Network of Excellence, 6th Framework Programme, Information Society Technology, supported by OFES

Duration: May 2004 – April 2007

Partners: 121 institutes, all European countries

Contact persons: Joé Millan

Description: EURON focuses on robotics technologies "beyond" the factory floor; e.g., to assist people in their daily life and to augment their capabilities beyond natural boundaries.

Europe is already today the leader in industrial robotics. At the same time Europe is experiencinga significant aging of society. This change in demographics will have consequences on industry, style of living, entertainment, etc. A key contributor to the development of aids for everyday life (at the workplace and in the homes) will be robotics technology. The topic is, however, far wider than traditional industrial robotics, it involved direct brain interfaces, service robotics, etc. In addition, the area is not only in need of new research and development, but also human resources to participate and drive the innovation process. To ensure that the economic growth in robotics remains in Europe, there is a need to unite the R&D, teaching and dissemination activities across the entire union and associated states. The work-plan involves activities on research coordination (across the pro-active initiative), training and education, dissemination efforts, collaboration with end-user industries, and research efforts on emerging problems.

FGnet – Face and Gesture Recognition Working Group

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

Duration: 36 months, September 2001-August 2004

Partners: University of Manchester, Gerhard-Mercator-University Duisburg, Aalborg University, Institut National Polytechnique de Grenoble, Cyprus College

Contact person: Sebastien Marcel

Description: FGnet is a "Concerted Action and Thematic Network" on Face and Gesture Recognition. The use of shared resources and data sets to encourage the development of complex process and recognition systems has been very successful in the speech analysis and recognition field, and in the image analysis field in the specific cases where it has been applied. The aim of the project is thus to encourage the development of common databases, technological approaches, and evaluation standards in the area of face and gesture recognition, i.e.:

1. Providing focus and common grounds for researchers developing face and gesture recognition technology
2. Creating a set of foresight reports defining development roadmaps and future use scenarios for the technology in the medium (5-7 years) and long (10-20 years) term
3. Specifying, developing and supplying resources (e.g. image sets) supporting these scenarios. The resource generation activity will involve the specification of key data sets, evaluation protocols and reference architectures that will form the basis for technology development and sharing.
4. Encouraging the use of these resources to share and boost technology development.

GHOST – Gesturing Hand recOgition baSed on user Tracking

Funding: France Telecom R&D

Duration: 24 months, February 2002-February 2004

Partners: Telecommunication and Neural Techniques group of France Telecom R&D DTL/TIC

Contact person: Sébastien Marcel

Description: The aim of the project is thus to recognize up to 15-20 hand gestures. It is necessary to distinguish two aspects of hand gestures :

- the static aspect is, for instance, characterized by a posture of the hand in an image,
- the dynamic aspect is defined either by the trajectory of the hand, or by the sequence of hand postures in a sequence of images.

In this project hand gestures are represented by trajectories of the hand in 3D. The hand gesture database is provided by France Telecom R&D and is acquired using a stereo camera framework. Our work has as an ambition to develop hybrid techniques of statistical training (Hidden Markov Models and Neural Networks) for the recognition of hand gestures (pointing gestures or drawing gestures).

HOARSE – Hearing Organisation and Recognition of Speech in Europe

Funding: European project, 5th Framework Programme, Training and Mobility of Researchers (TMR) programme, Research Network, supported by OFES

Duration: 48 months, September 2002-August 2006

Partners: Sheffield University (UK), Rurh- University Bochum (D), Daimler-Chrysler (D), Helsinki University (FIN), Keele University (UK), Patras University (G), IDIAP (CH)

Contact person: Hervé Bourlard

Description: As a follow-up of the SPHEAR TMR project (see below), the overall objectives of HOARSE are to gain a better understanding of speech production and hearing mechanisms and to use this understanding to explain the perceptual organization of sound and improve speech technology. This project will thus involve several research themes, including:

1. Auditory Scene Analysis: Understanding how sound mixtures are perceptually organized into a coherent auditory scene, and how this organization can be used in speech recognition.
2. Dealing with Reverberant Conditions: Reverberant conditions are a big problem for speech recognition, and their processing in human hearing.
3. Speech Production Modelling: Understanding how speech is produced, how this relates to speech perception and cerebral speech processing, and how this knowledge can be integrated in state-of-the-art speech recognition systems.
4. Automatic Speech Recognition Methodologies: Generalization of state-of-the-art automatic speech recognition algorithms to take advantage of the above. Specifically, we focus here on natural listening conditions, where the speech to be recognized is one of many sound sources (including noise and competing speech) which change unpredictably in space and time.

HMM2 – A New Framework for Robust and Adaptive Speech Recognition

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Ikbal Shajith, Hervé Bourlard

Description: The HMM2 project is directed towards extending the hidden Markov model (HMM) framework to simultaneously accommodate complex constraints in both the temporal and frequency domains. The generic idea of the approach investigated here, referred to as HMM2 for obvious reasons, is to associate with each (temporal) HMM-state a second, frequency based, HMM which will model the underlying probability density function. In other words, the multi-gaussians (or artificial neural network) typically used in standard HMMs will be replaced by a frequency-based HMM, responsible for estimating, through frequency-based latent variables, the "temporal" HMM emission probabilities and the correlation across the frequency bands.

Such an approach (for which standard multi-gaussians are a particular case) has many potential advantages, including: (1) in the case of multi-band speech recognition, dynamic definition and adaptation of the subbands, (2) automatic formant tracking, (3) nonlinear frequency warping, and (4) modeling of the correlation across frequency bands.

IM2.ACP - Access and Content Protection

Funding: SNSF, through the (IM)2 NCCR

Duration: January 2002 – December 2003

Contact persons: Norman Poh, Conrad Sanderson, Samy Bengio, Hervé Bourlard

Description: In the framework of IM2.ACP, IDIAP is mainly developing new text-dependent (user-customized) and text-independent speaker verification systems. IDIAP is also investigating advanced multimodal biometric identification/verification systems, trypically based on voice and face verification, and involving different fusion algorithms.

Experiments on scalability of speaker verification and fusion algorithms have been performed in order to verify how the system degrades when the size of the model needs to be small. The results are very interesting:

- the performance of the speaker verification system alone degrades slowly with respect to the reduction of the number of parameters (logarithmically at the beginning, hence we can remove a lot of parameters without a big performance loss)
- the performance of the fusion system degrades even more slowly than the speaker verification system, since the degradation of unimodality systems (speaker and face systems) are independent from each other.

These experiments were carried out on the English BANCA database and its associated protocol. Further experiments on variations on the creation and use of World Models for text independent speaker verification have shown promising results with respect to cross-gender attacks. These experiments were carried out on the NIST and PolyVar databases and novel protocols were designed to verify specific attacks.

IM2.IP - Integration project

Funding: SNSF, through the (IM)2 NCCR

Duration: January 2002 – December 2005

Contact persons: Pierre Wellner, Hervé Bourlard

Description: The general goal of the IM2 Integration Project is to leverage on the research and development efforts available within IM2. In that framework, IM2.IP is also responsible for consolidating as much as possible all of the IM2 outcomes in common projects and applications. More specifically, IM2.IP is focusing on:

- Smart Meeting Room (SMR)
- Collection, annotation, and distribution of data
- Extension and integration of new SMR functionalities
- Integration of multimodal technologies

IM2.MI - Multimodal Integration

Funding: SNSF, through the (IM)2 NCCR

Duration: January 2002 – December 2003

Contact persons: Mikaela Keller, Samy Bengio, Hervé Bourlard

Description: The goals of IM2.MI are the research and development of principled methods for the fusion and efficient decoding of different input modalities (multi-channel processing). The objectives can thus be decomposed as follows:

- Development of new multi-channel, multi-rate, signal processing techniques (including EEG processing)
- Development of new data fusion algorithms, as well as new decision strategies
- Development of a multichannel statistical model for the combination of asynchronous input streams
- Development of an efficient multimodal decoder
- Implementation of all these algorithms into a common software platform.

During the first year of this project, IDIAP started working on a new HMM model to handle asynchronous streams of data. This model is mainly inspired by two other models, namely:

- Asynchronous Input/Output Hidden Markov Models (AIOHMM), which enables modelling of an output sequence conditioned on an input sequence asynchronously, and
- Multi-stream HMMs.

In a recent paper (part of a special ICASSP session; see also IDIAP Research Report 02-59), the above approach has been used to analyse multimodal meeting behaviors. In this case, multiple streams containing audio and video information related to all participants of a meeting were merged in order to decode the general behavior of the meeting. The possible behaviors follow a language model with such events as monologues, discussions, presentations, consensus, disagreements, etc. The overall objective of the project is to summarize a new meeting according to this language. Several small meetings were recorded in order to train the model. Preliminary experiments show that such a multimodal decoder can obtain up to 80% event recognition on new meetings involving persons that were not present in any of the training meetings. Further experiments involving asynchronous HMMs will be performed soon.

IM2.BMI - Brain Machine Interfaces

Funding: SNSF, through the (IM)2 NCCR

Duration: July 2002 – June 2004

Contact persons: Silvia Chiappa, Josè del R. Millàn, Hervé Bourlard

Description: In the framework of IM2, a project on Brain Machine Interface has also been defined. Addressing a particular kind of (advanced) man-machine interface, this project aims to investigate the possibility of classifying spontaneous brain activity based on either reconstructed brain activity maps, or directly from EEG recordings (thus, a particular kind of multi-channel processing).

During the first few months of this project, the following has been achieved:

- A communication infrastructure was created in order to facilitate the relations between the partners of the consortium. An email list as well as a protected website were created and are now maintained at IDIAP.
- The experimental system has been set up (see Figure 10), and an initial set of data has been recorded and is currently being analyzed by the partners.
- State-of-the-art survey: At least 20 identified research groups are working on Brain- Computer/Machine Interfaces (BCI or BMI). Currently, it is possible to classify from 2 to 10 different mental states, within a precision ranging from 80% to 100%. This means a bit rate of 1 to 100 bits/min, which is 10 to 100 times lower than keyboard bit rate.

IM2.MUCATAR - Multiple Camera Tracking and Activity Recognition

Funding: SNSF, through the (IM)2 NCCR

Duration: July 2002 – June 2004

Contact persons: Sileye Ba, Kevin Smith, Jean-Marc Odobez, Daniel Gatica-Perez

Description: In the context of IM2, human tracking and event/activity recognition represent essential components towards multimodal interaction and analysis of multimedia databases. The understanding of human activities in indoor environments is important, both as a component of a multimodal interface, and to extract semantic clues from videos for indexing and retrieval. The main research problems that will be addressed in the project are :

- the development of exemplar-based models of typical people activity that can be constructed directly from training sets in a probabilistic setting;
- the combination of multiple visual features for tracking;
- the study of the combination of Sequential Monte Carlo and Hidden Markov Models for performing jointly the tracking and recognition tasks;
- the extension of such formulations to a multiple camera scenario.

The project started with the study and the implementation of standard particle filters (PF), with shape and color object models. After, we explored the use of Importance PF to perform the assymmetrical integration of audio-visual information in a way that efficiently exploits the complementary features of each modality. This research has been applied to the audio-visual tracking of multiple speakers in meeting rooms, as illustrated in Figure 8. At the same time, a new probabilistic model for visual tracking has been proposed. This model allows for an implicitly modeling of motion, and early results show that this model leads to more stable and discriminative trackers than those using generic object modeling alone (e.g. with shape and color).

IM2.RTMAP - Reat Time Microphone Array Processing

Funding: SNSF, through the (IM)2 NCCR

Duration: July 2002 – June 2004

Contact persons: Olivier Masson, Darren Moore, Iain McCowan, Hervé Bourlard

Description: An IM2 white paper project, entitled "Real-time Microphone Array Processing for Meeting Room" has commenced within the framework of IM2.SP. The project aims to use microphone arrays to address the problems of acquiring clean speech, detecting periods of voice activity and dynamically determining the location for each meeting participant. The major objectives are to research novel array processing techniques aimed at making arrays more viable in terms of cost, processing and space requirements, and also to produce a stand-alone system that will facilitate further research in IM2. In this framework, and as illustrated in Figure 11, new microphone array algorithms have been developed using new post-filter formulated for diffuse noise field (see IDIAP RR-39, to be published in IEEE Trans on Signal Processing).

Within the scope of the white paper, a sub-project & contract to HEV-Sion has been defined, entitled "Small Microphone Array Low Level Development Project". In this sub-project, IDIAP has subcontracted to HEV Sion the task of implementing the hardware and low-level software required for the proposed modular array architecture. The sub-project commenced in August under the direction of Dr. Joseph Moerschell and has an expected duration of 6 months. The major deliverable for the sub-project is the demonstration of a stand-alone 8 channel microphone array with basic functionality.

IM2.SP - Speech Processing

Funding: SNSF, through the (IM)2 NCCR

Duration: January 2002 – December 2003

Contact persons: Hemant Misra, Hervé Bourlard

Description: The goal of IM2.SP is to provide the IM2 NCCR with advanced and flexible speech processing modules which can be used as an input mode (voice input), as an audio indexing tool (requiring large vocabulary, continuous speech recognition systems) turning audio files into text, and as an output mode (mainly based on text-to-speech systems). The research goals of this IP are thus to improve state-of-the-art speech recognition algorithms (with respect to performance and robustness to noise and speech style), as well as (to a lesser extent) text-to-speech technologies.

During the first 9 months of IM2, and following the planned activites briefly presented above, the following has been achieved:

- Development of the TODE Decoder:

  The TODE recogniser developed at IDIAP is designed to meet the speech decoding needs of researchers at IDIAP and in the wider speech research community, and to insure easy adaptation and porting to different tasks and environments. In its current implementation, the main functionalities and features of TODE are:

  - Based on a time synchronous Beam Search decoding algorithm
  - Integrated with the TORCH machine learning toolkit
  - Accepts feature vectors or acoustic probability vectors as input
  - Supports both GMM and ANN-based acoustic modelling

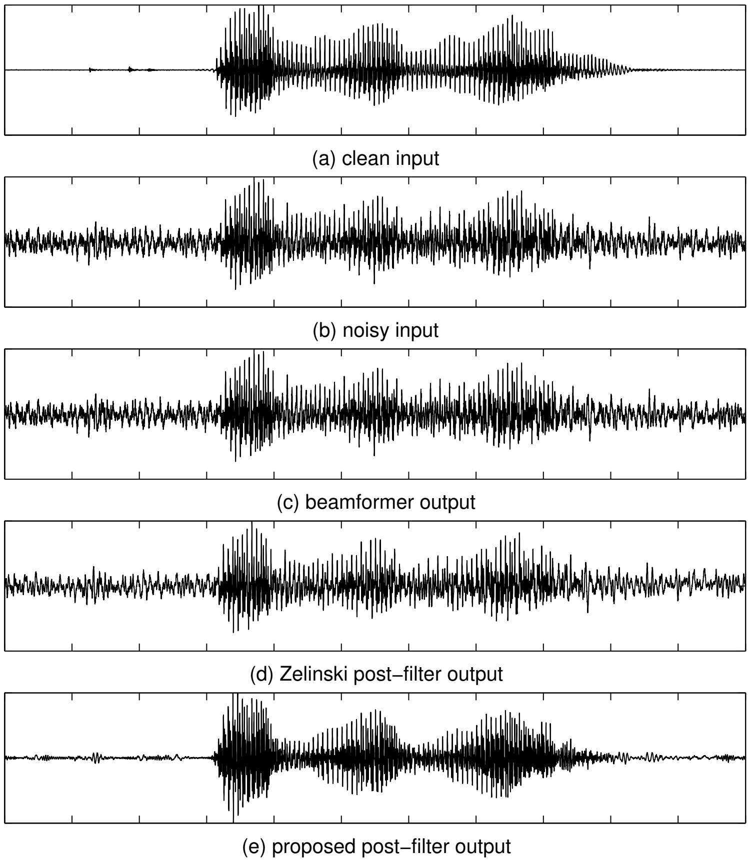

Figure 11: Plots demonstrating speech enhancement using a 5 element microphone array. (a) original clean input signal; (b) noise corrupted input signal; (c) output of standard superdirective beamformer; (d) output of standard array post-filter; (e) output of proposed IDIAP post-filter optimised for diffuse noise field.

  - Arbitrary N-gram language modelling
  - Compatible with many popular file formats
  - Linear lexicon
  - Integrated Word Error Rate (WER) calculation
  - Supported - development ongoing

  Manual of the speech decoder can now be downloaded from http://www.torch.ch/documentation.php.

- New algorithms for automatic audio segmentation have been developed and tested, including:
  - A particularly performant system for speech/non-speech detection has been developed and tested on several international databases.
  - A new approach towards automatic speaker clustering and speaker turn detection has been developed and tested. This system is based on a new information theory based clustering (generalizing the BIC criterion originally proposed by IBM) determining the optimal number of clusters without the need of any penalty term (as opposed to BIC).
- Work on microphone array (IDIAP) recording and sound source localization has just started. In this framework, one new PhD student (Guillame Lathoud) has been hired to work on this. See Section 2.2.1 for further detail.

Kernel Methods for Sequence Processing

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Quan Le, Samy Bengio

Description: *Hidden Markov Models* (HMMs) are one of the most powerful statistical tools developed in the last twenty years to model sequences of data such as time series, speech signals or biological sequences. One of their distinctive features lies on the fact that they can handle sequences of varying sizes, through the use of an internal state variable.

Unfortunately, it is well known that for classification problems, a better solution should in theory be to use a *discriminant* framework. In that case, instead of constructing a model independently for each class, one constructs a unique model that decides where the frontiers between classes are.

A series of recent papers have suggested some possible techniques that could be used to mix generative models such as HMMs (to handle the sequential aspects) and discriminant models such as Support Vector Machines.

The purpose of the present project is thus to study, experiment (on different kinds of sequential data), enhance, and adapt these new approaches of integrating discriminant models such as SVMs into generative models for sequence processing such as HMMs.

LAVA – Learning for Adaptable Visual Assistants

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

Duration: 36 months, March 2002-February 2005

Partners: Xerox Research Center Europe (UK and France), INRIA (F), University of London (UK), Lund University (S), Graz University of Technology (A), IDIAP (CH), Australian National University (AUS)

Contact person: Samy Bengio, Jean-Marc Odobez, Mark Barnard, Pedro Quelhas, Alexei Pozdnoukhov

Description: The overall objective of LAVA is to create fundamental enabling technologies for cognitive vision systems. The resulting widely transferable knowledge is to be thoroughly evaluated and widely disseminated. The new technologies that LAVA will provide will enable new tools for a wide range of applications including "ambient intelligence scenarios". The project includes the following objectives:

- Robust and efficient categorisation and interpretation of large numbers of objects, scenes and events, in real settings
- Automatic acquisition of knowledge of categories, for convenient construction or extension of applications.

MAIA – Non-Invasive Brain Interaction with Robots – Mental Augmentation through Determination of Intented Action

Funding: EU STREP IST project, 6th FWP

Duration: June 2004 - May 2007

Partners: IDIAP (coordinator), Katholieke Universiteit Leuven (B), University Hospital of Geneva (CH), Fondazione Santa Lucia-Rome (I), Helsinki University of Technology (F)

Contact persons: Joé Millan

Description: MAIA aims at developing non-invasive prosthesis. In particular, a brain-computer interface recognizes the subject's voluntary intent to do primitive motor actions on the order of milliseconds and conveys this intention to a robot that implements the necessary low-level details for achieving complex tasks. To achieve this objective, we will take a radical departure from current assumptions and approaches in BCI. In particular, we will follow five innovative principles: - recognition of the subject's motor intent from the analysis of high resolution brain maps, which estimates intracranial potentials from scalp EEG, what would facilitate scaling up the number of mental commands and increasing recognition speed; - adaptive shared autonomy between two intelligent agents–the human user and the robot–so that the user only gives high-level mental commands that the robot performs autonomously; - use of haptic feedback to the user to accelerate training and facilitate accurate control; - recognition of brain events associated to high-level cognitive states, such as error recognition and alarm, to increase the reliability of the brain-actuated robots; - on-line adaptation of the interface to the user to keep the BCI constantly tuned to its owner.

These principles will be demonstrated in three applications, which will be used to measure the S&T objectives of the MAIA project. The three demonstrations are: - driving a wheelchair in an indoor environment; - controlling a robot arm for reaching and manipulation tasks; - handling emergency situations after recognizing the subject's alarm state (e.g., braking the vehicle or retracting the robot arm).

Noisy Text Retrieval

Funding: Swiss National Science Foundation, now part of the global MULI project

Duration: see MULTI

Contact persons: David Grangier, Alessandro Vinciarelli

Description: Several media contain textual information that can be accessed only through an extraction process (e.g. the text contained in spoken documents or handwritten texts can be extracted through a recognition process).

The extraction process produces *noise*, i.e. the extracted text is different from the actual clean text contained in the original media: the noisy text contains word substitutions, insertions and deletions.

Information Retrieval (IR) aims at identifying the documents relevant to a query in a text collection. Researches in IR focused essentially on clean text collections. The extension of IR techniques to noisy texts allows one to retrieve data contained in other media sources such as audio and video recordings.

For example, a database of speech recordings can first be converted into digital texts using an Automatic Speech Recognition system. The transcriptions can then be searched with an IR system.

The information contained in audio recordings is not limited to textual data. Information such as speaker identity or dialog dynamics can be extracted. The use of such information can enhance the retrieval process.

M4 – MultiModal Meeting Manager

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

Duration: 36 months, March 2002-February 2005

Partners: Sheffield University (UK), München University (D), TNO/TPD (NL), University of Twente (I), EPFL/LTS (CH), UniGe (CH), IDIAP (CH), ICSI (Berkeley, CA).

Contact person: Hervé Bourlard, Daniel Gatica-Perez

Description: The overall aim of M4 is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings. The archived meetings will have taken place in a room equipped with multimodal sensors. For each meeting, audio, video, textual, and (possibly) interaction information will be available. Audio information will come from close talking and distant microphones, as well as binaural recordings. Video information will come from multiple cameras. While the video and audio information will form several streams of data generated during the meeting, the textual information (the agenda, discussion papers, text of slides) will be pre-generated and will be used to guide the automatic structuring of the meeting. The interaction stream consists of any information that can help in analysing events within the meeting, for example, mouse tracking from a PC-based presentation or laser pointing information. The main research and development streams of M4 thus include::

1. Development of a "smart" meeting room, collection and annotation of a multimodal meetings database.
2. Automatic analysis and processing of the audio and video streams, including: robust conversational speech recognition, recognition of gestures and actions, multimodal identification of intent and emotion, multimodal person identification, source localization and tracking.
3. Integration and structuring using the output of the various recognizers and analyses, including: specification of a flexible intelligent information management framework, models for the integration of multimodal stream, summarization of a meeting, or a meeting segment, multimodal information extraction and cross-lingual retrieval/browsing across the archive.
4. Construction of a demonstrator system for browsing and accessing information from an archive of processed meetings.

Multimodal Interaction and Multimedia Data Mining

Funding: Swiss National Science Foundation

Duration: October 2002 - September 2004

Contact persons: Hervé Bourlard

Description: Since October 1, 2002, this new NSF project is integrating (and extending) all the SNSF research activities (apart from the IM2-NCCR activities) reported in the present document. As a unified research theme, the goal of the present Swiss National Science Foundation (SNSF) project will be to carry out fundamental research in the field of multimodal interaction, which covers a wide range of critical activities and applications, including recognition and interpretation of spoken, written and gestural language. Other key subthemes of this project will include the control of information access, typically through biometric user authentication techniques (including speaker and face verification). Building upon the same technologies, the present project will also investigate advanced approaches towards the structuring, retrieval and presentation of multimedia information, also referred to as "multimedia data mining".

This is indeed a wide-ranging and important research area that includes not only the multimodal interaction described above, but also multimedia document analysis, indexing, and information retrieval, thus involving complex computer vision and data fusion algorithms.

FNSNF Optimization for Machine Learning

Funding: Swiss National Science Foundation, now part of the global MULI project

Duration: see MULTI

Contact persons: Ronand Collobert, Samy Bengio

Description: As organizations collect more and more data, the interest in extracting useful information from these data sets with *data mining* algorithms is pushing much research effort toward the challenges that these data sets bring to statistical learning methods. One of these challenges is the sheer size of the data sets: many learning algorithms require training time that grows too fast with respect to the number of training examples. This is for example the case with Support Vector Machines (SVMs) a non-parametric learning method that can be applied to classification, and regression. They require $O(T^3)$ training time (for T examples) in the worst case or with a poor implementation. Empirical computation time measurements on state-of-the-art SVMs implementations show that training time grows much closer to $O(T^2)$ than $O(T^3)$. It has also been conjectured that training of Multi-Layer Perceptrons (MLPs) might also scale between quadratic and cubic with the number of examples.

It would therefore be extremely useful to have general-purpose algorithms which allow to decompose the learning problem in such a way as to drastically reduce the training time, so that it grows closer to $O(T)$. It would be also useful if one could improve standard training algorithms for SVMs and MLPs.

FNSNF Multimodal Group Action Recognition

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Dong Zhang, Daniel Gatica-Perez

Description: The problem addressed here is the analysis of multimodal information in the context of meetings. Meetings play an important role in everyday life. Therefore, it is relevant to develop models and algorithms to enable efficient access to archived meetings. Meetings take place in the IDIAP smart meeting room, equipped with computers and multiple sensors.

Meetings differ from many other multimodal processing tasks due to their group nature, as a result of the interactions between participants. In our view, multimodal group action recognition is one of the ultimate goals of automatic meeting analysis, as it can summarize a meeting into a sequence of high-level items. Though speech is the predominant modality, meetings are multimodal in nature, as described in existing literature in social psychology. Therefore, meetings provide a ideal platform for multimodal research.

In this work, we propose to use machine learning techniques (mainly generative sequence models) and multi-modal cues to recognize both individual actions and group actions with applications to meeting analysis. Up to this moment, I have proposed to extend the ideas originally developed at IDIAP, dividing the group action recognition problem into two layers. The lower layer is performed using standard probabilistic models such as hidden Markov models (HMMs) to recognize actions of individual participants. The input of this layer are low-level audio-video features extracted from each person. The resulting individual action models are intended to be person-independent, since the training data come from different persons. The output of the lower level (individual action) provides the input to the second layer of the framework, which recognizes group actions from individual actions and a set of group events.

PASCAL – Pattern Analysis, Statistical Modelling and Computational Learning

Funding: European Network of Excellence, 6th Framework Programme, Information Society Technology, supported by OFES

Duration: January 2004 – December 2007

Partners: 57 institutes, 17 countries, around 120 researchers; IDIAP is the Swiss Coordinator and official EC link with AMI Integrated Project.

Contact persons: Samy Bengio, José Millan

Description: The objective of the PASCAL network is to build a Europe-wide Distributed Institute which will pioneer principled methods of pattern analysis, statistical modelling and computational learning as core enabling technologies for multimodal interfaces that are capable of natural and seamless interaction with and among individual human users.

At each stage in the process, machine learning has a crucial role to play. It is proving an increasingly important tool in Machine Vision, Speech, Haptics, Brain Computer Interfaces, Information Extraction and Natural Language Processing; it provides a uniform methodology for multimodal integration; it is an invaluable tool in information extraction; while on-line learning provides the techniques needed for adaptively modelling the requirements of individual users. Though machine learning has such potential to improve the quality of multimodal interfaces, significant advances are needed, in both the fundamental techniques and their tailoring to the various aspects of the applications, before this vision can become a reality.

The institute will foster interaction between groups working on fundamental analysis including statisticians and learning theorists; algorithms groups including members of the non-linear programming community; and groups in machine vision, speech, haptics, brain-computer interfaces, natural language processing, information-retrieval, textual information processing and user modelling for computer human interaction, groups that will act as bridges to the application domains and end-users.

PRIOR Knowledge in Kernel Methods

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Alexei Pozdnoukhov, Samy Bengio

Description: A number of contemporary Machine Learning algorithms known as *kernel methods*, including Support Vector Machines, are based on the idea of the "kernel trick". Given a proper *kernel function* that defines the dot product in a feature space and a ML algorithm formulated in terms of the dot products between the input samples, one easily obtains a non-linear form of the algorithm by replacing the dot products with the kernel function. The application field of kernel algorithms (basically, SVMs) is quite wide and the reported results are promising.

An important question which arises in this framework is *the choice of the kernel function* as it explicitly defines the feature space hence it is of crucial meaning for the performance of the algorithms. The kernel function is also a factor that defines the capacity of the model. *Kernel design methodology*, and the particular problem of *incorporating some prior knowledge into the kernel function* is the mainstream of the research project. The incorporation of the prior knowledge on the *invariances* of the problem is of our main interest.

The research is mostly oriented to obtaining a general approach to the problem rather then in particular application-dependent solutions. So far, the general "From Sample to Object" approach was formulated and applied for the stated problem.

PROMO – PROnunciation MOdelling in Automatic Speech Recognition Systems

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Mathew Magimai Doss, Hervé Bourlard

Description: Natural speech and casual human conversation exhibit a large amount of nonstandard variability in pronunciation. Phonological studies of the way a word is pronounced in different lexical contexts by native speakers of a language in clearly articulated speech lead to more than one acceptable pronunciation for many words. This results in a mismatch between the baseline phonetic transcriptions given in the lexicon and the actual pronunciation of the words, seriously hindering the recognition performance.

The mismatch between the dictionary representation of words and their actual realization may be reduced using an improved pronunciation model. In state-of-the-art speech recognition systems, this is often achieved simply by adding many pronunciation alternatives for each word, or by automatically inferring pronunciation variants from multiple utterances of each word.

The main motivation of this project is thus to investigate new techniques towards robust modelling of pronunciation variants in the context of continuous speech recognition, and more particularly in the case of natural speech recognition. On top of further investigating standard approaches (such as the automatic generation of pronunciation variants based on a maximum likelihood criterion), this project will focus on (1) dynamic pronunciation modelling, and (2) discriminant training of pronunciation models.

Semantic Indexing of Multimedia Documents with Generative Probabilistic Models

Funding: Swiss National Science Foundation, now part of the global MULI project

Duration: see MULTI

Contact persons: Florent Monay, Danial Gatica-Perez

Description: Users that access *large image collections* always face the same problem. Because of the amount of images, they need ways to search through them; however,for the same reason, they cannot expect the images to be fully annotated by hand, as this represents a huge and expensive job.

A basic solution consists in providing a *query image* that roughly looks like what a user is looking for, then computing the similarity between the visual features (colors, spatial frequencies, . . .) of the query and all the images in the collection. After a long iterative query refinement process called *relevance feedback*, a user might end up with some images that hopefully correspond to what he/she originally wanted[1]. While this process sometimes leads to good results (depending on the query and the datatset complexity), this retrieval process is somewhat contradictory: the user already has to have the image he/she is looking for.

Searching an image collection in a *semantic* manner - using words as queries - is more intuitive, and highly attractive, as exemplified from the success in web text-based search/retrieval. But, as described before, the annotation of images (ranging from only filenames to full text captions) is usually sparse. One

[1] see demo at http://viper.unige.ch/demo/php/demo.php

attractive option is to use machine learning techniques, namely probabilistic generative models, to learn the joint distribution of visual features and words. Doing so provides several automatic capabilities that improve the retrieval process: annotation of non-labelled images (auto-annotation[2]) and of image regions (object recognition), enrichment of existing image annotations (annotation propagation), and illustration of word-based queries (auto-illustration).Furthermore, it represents one of the most promising attempts to bridge the gap between low-level visual features and semantic concepts. The development of statistical models for the joint modelling of textual and visual information is a very active research area, and the central goal of this project.

FN SNF SCRIPT – Cursive Handwriting Recognition

Funding: Swiss National Science Foundation, now part of the MULTI project

Duration: see MULTI

Contact person: Alessandro Vinciarelli

Description: The recognition of cursive handwritten words when only the image of the data is available is called Off-Line Cursive Script Recognition (CSR). The great variability of handwriting styles and the fact that the letters are connected are the major difficulties of the problem.
A system for single word recognition was developed. It presents an original normalization method (based on statistics) that improved significantly the performance with respect to traditional normalization methods.
We are extending now the recognition problem to the automatic reading of sentences.

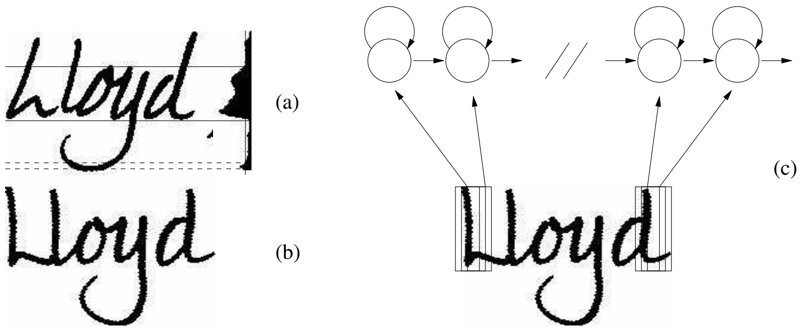

Figure 12: Single word recognition. The original image (a) is normalized (b) and modelled with HMMs (c). A HMM is created for every word in a list of possible interpretations of the data. The most likely model is assumed as transcription of the data.

Language modelling Unlike the case of single word recognition, it is possible to apply language modelling techniques to improve the performance. The n-gram models, the current state of the art, will be extensively applied in order to verify their effectiveness in the handwriting problem. Furthermore, language models only partially successful in speech domains (i.e. stochastic grammars), can be probably more helpful when applied to the written communication that is, in general, more formal than the oral one.

[2]See examples of auto-annotation at http://www.idiap.ch/∼monay/AnnotExRandom.php

Search Technique The recognition of the handwritten data consists in measuring the matching between the observations (the vectors extracted from the data images) and the sentence models (HMM concatenations). This is done by finding the optimal path (in terms of some specified criterion) in a properly structured search space. This must involve both local (single letter level) and global (language model level) constraints. Besides, pruning techniques must be studied and applied in order to limit as much as possible the number of hypotheses considered (without reducing the overall recognition performance).

Hidden Markov Modelling Several parameters require to be set in Hidden Markov Models: number of states, topology, number of Gaussians in mixtures. Accurate experiments will be performed in order to find their optimal values. Moreover, an approach successfully applied in speech recognition will be applied, the hybrid HMM/ANN architecture.

SV-UCP – Speaker Verification based on User-Customized Password

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Mohamed Benzeghiba, Hervé Bourlard

Description: The general objective of the present project is to further improve state-of-the-art speaker verification systems, where IDIAP has a recognized leading position. More specifically, the aim of this project is to investigate new alternatives to speaker verification systems, based on user-customized password (allowing the user to choose his/her password, just by pronouncing it a few times).

In the context of this project, automatic HMM inference approaches and fast speaker adaptation techniques will be investigated. This research is carried out in the framework of standard HMM, as well as in the context of hybrid HMM/ANN systems. Particular attention is however paid to the use of HMM/ANN systems since ANN have been shown to yield significantly better phonetic classification performance, which should potentially benefit to the precision of the automatically inferred HMMs (from a few pronunications of the password). On the basis of that inferred HMM, different speaker adaptation techniques are also being studied, and the resulting speaker verification performance is assessed on the Polyvar reference database.

VOCR - Text Recognition for Video Retrieval

Funding: Swiss National Science Foundation, now part of the global MULTI project

Duration: see MULTI

Contact persons: Datong Chen, Jean-Marc Odobez

Description: 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.

This year, the research has focused on three main issues :

- the design of a fast text localization focusing step, which enables text size normalization. It relies on a machine learning text verification step applied on background independent features.
- the improvement of text recognition results. It is addressed by a text segmentation step followed by an traditional OCR algorithm within a multi-hypotheses framework relying on multiple segments, language modeling and OCR statistics.
- the exploitation of temporal information to reach a better decision, correct errors. A selection mechanism of the best solution over time has been designed. Cuurrently, we are studying the possibility of merging solutions to produce a new recognition string.

All these techniques have been implemented and incorporated into software to be used in the european projects ASSAVID and CIMWOS. Experiments conducted on large databases of real broadcast documents provided by the partners (BBC, Canal+) have proven the robustness of our approach.

KTI CTI VoiceInPack – Low bit-rate speech transmission based on speech recognition and speech synthesis for online multiplayer games

Funding: Swiss Commission for Technology and Innovation (CTI)

Duration: November 2002 - April 2004

Partners: ETHZ and Komodo Entertainment Software SàRL

Contact persons: Hervé Bourlard

Description: VoiceInPack will contribute to the development of a Massively Online Game created by Komodo. VoiceInPack will develop a technology that allows, with a high level of compression, to fullfill Komodo's game voice requirements. VoiceInPack aims at developing very low bit rate speech transmission (over the internet, for game applications) by using the front-end of a speech recognition system to turn the voice signal into a sequence a phone (developed by IDIAP), complemented by additional prosodic parameters (such as pitch and duration). These phoneme sequences will then be sent over the communication channel, and later used as the input of the back-end of a text-to-speech system (developed by ETHZ). The output signal will also be modulated accordingly to the player's avatar. While interesting from an application point of view, this project will also allow further research and development activities in speech recognition (improvement of phonetic recognition) and speech synthesis (using prosodic features to improve naturalness).

7 Educational Activities

7.1 Current PhD Theses

The list of current IDIAP PhD students, together with their PhD projects and funding sources, is summarized in the table on next page. For a brief description of their research projects, please refer to Section 6.

7.2 IDIAP PhD Defenses

- **Ph.D. candidate:** Datong Chen

  Supervisor: Jean-Marc Odobez - Hervé Bourlard

  Examiners: J. Kittler - J.M. Jolion - J.Ph. Thiran - R. Hersch - J.M. Odobez - H. Bourlard

  University: EPFL

  Date: 28 October 2003

  Title: Text Detection and Recognition in Images and Video Sequences

- **Ph.D. candidate:** Todd Stephenson

  Supervisor: Hervé Bourlard

  Examiners: F. Rachidi - S. King - G. Zweig - M. Hasler - S. Bengio

  University: EPFL

  Date: 28 May 2003

  Title: Causal Based Automatic Speech Recognition with Dynamic Bayesian Networks

- **Ph.D. candidate:** Alessandro Vinciarelli

  Supervisor: Samy Bengio

  Examiners: H. Bunke - H. Bourlard - O. Nierstrasz

  University: Bern

  Date: 2 April 2003

  Title: Offline Cursive Handwritting: From Word to Text Recognition

- **Ph.D. candidate:** Katrin Weber

  Supervisor: Hervé Bourlard

  Examiners: J.R. Mosig - S. Bengio - A. Billard - M. Cooke - G. Rigoll

  University: EPFL

  Date: 28 May 2003

  Title: HMM Mixtures (HMM2) for Robust Speech Recognition and Formant Tracking

PhD Students		Expected PhD	At IDIAP since	PhD Status	PhD Area	IDIAP Thesis Supervisor(s)	Thesis Director
Ajmera	Jitendra	2004	01.01.01	4th year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Ba	Silèye	2006	01.10.02	2nd year	Computer Vision	J.M. Odobez + S. Bengio	Not decided yet
Barnard	Marc	2005	15.03.01	3rd year	Computer Vision	J.M. Odobez + S. Bengio	Prof. H. Bourlard, EPFL
Benzeghiba	Mohammed F.	2004	01.08.00	4th year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Cardinaux	Fabien	2005	01.10.01	3rd year	Computer Vision	S. Marcel	Prof. J.-Ph. Thiran, EPFL
Chen	Datong		01.11.99	Accepted	Computer Vision	J.M. Odobez	Prof. J.-Ph. Thiran, EPFL
Chiappa	Silvia	2005	01.11.01	3rd year	Machine Learning	S. Bengio + J. Millan	Prof. H. Bourlard, EPFL
Collobert	Ronan	2004	01.08.02	4th year	Machine Learning	S. Bengio	Prof. P. Gallinari, Univ. Pierre et Marie Curie, Paris
Dimitrakakis	Christos	2005	01.10.01	3nd year	Machine Learning	S. Bengio	Prof. H. Bourlard, EPFL
Fasel	Beat	2004	01.10.98	Completed	Computer Vision	D. Gatica-Perez	Prof. Van Gool, ETHZ
Grangier	David	2007	01.10.03	1st year	Speech Processing	A. Vinciarelli	Prof. H. Bourlard, EPFL
Ikbal	Shajith	2004	01.05.00	4th year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Just	Agnès	2006	01.10.02	2nd year	Computer Vision	S. Marcel + S. Bengio	Prof. H. Bourlard, EPFL
Keller	Mikaela	2006	01.12.02	2nd year	Machine Learning	S. Bengio	Prof. H. Bourlard, EPFL
Lathoud	Guillaume	2006	01.03.02	2nd year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Le	Quan	resigned	01.02.01				
Magimai Doss	Mathew	2004	25.10.99	4th year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Maier	Viktoria	2007	01.09.03	1st year	Speech Processing	H. Hermansky	Not decided yet
McGreevy	Michael	2007	15.01.03	2nd year	Speech Processing	H. Bourlard	Prof. Sridharan, QUT
Misra	Hemant	2005	24.07.01	3rd year	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Monay	Florent	2006	01.08.02	2nd year	Computer Vision	D. Gatica-Perez + S. Bengio	Prof. H. Bourlard, EPFL
Poh Hoon Thian	Norman	2006	01.09.02	2nd year	Machine Learning	S. Bengio	Prof. H. Bourlard, EPFL
Pozdnoukhov	Alexei	2006	01.07.02	2nd year	Machine Learning	S. Bengio	Prof. H. Bourlard, EPFL
Quelhas	Pedro	2006	01.11.02	2nd year	Computer Vision	J.M. Odobez + S. Bengio	Prof. H. Bourlard, EPFL
Rodriguez	Yann	2006	01.09.02	2nd year	Machine Learning	S. Marcel + S. Bengio	Prof. H. Bourlard, EPFL
Sivadas	Sunil	2005	01.09.03	1st year	Speech Processing	H. Hermansky	Prof. H. Bourlard, EPFL
Smith	Kevin	2006	21.11.02	2nd year	Computer Vision	D. Gatica-Perez + S. Bengio	Prof. H. Bourlard, EPFL
Stephenson	Todd		01.03.99	Accepted	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Tyagi	Vivek	resigned	01.06.01				
Vinciarelli	Alessandro		01.10.99	Accepted	Computer Vision	S. Bengio	Prof. H. Bunke, Uni Bern
Weber	Katrin		01.01.98	Accepted	Speech Processing	H. Bourlard	Prof. H. Bourlard, EPFL
Zhang	Dong	2007	01.08.03	1st year	Computer Vision	D. Gatica-Perez	Prof. H. Bourlard, EPFL

7.3 Participation in PhD Thesis Committees

- **Ph.D. candidate:** Torsten Butz
  Committee member: Samy Bengio
  University: EPFL
  Date: 10 June 2003
  Title: From Error Probability to Information Theoretic Signal and Image Processing
- **Ph.D. candidate:** F. de Wet
  Committee member: Hynek Hermansky
  University: Katholieke Universiteit Nijmegen, The Netherlands
  Date: 17 November 2003
  Title: Automatic speech recognition in adverse acoustic conditions
- **Ph.D. candidate:** Roberto Iglesias
  Committee member: José del R. Millán
  University: Univ. Santiago de Compostela
  Date: 10 January 2003
  Title: Nuevos Modelos de Cuantificación Vectorial basados en el Análisis Estructural del Conjunto de Datas
- **Ph.D. candidate:** Petr Motlicek
  Committee member: Hynek Hermansky
  University: Technical University Brno, Czech Republic
  Date: 25 November 2003
  Title: Modeling of Spectra and Temporal Trajectories in Speech Processingf
- **Ph.D. candidate:** Josep Mouriño
  Supervisor: José del R. Millán & Raimon Jané
  Examiners: Pere Caminal, Alícia Casals, Xavier Rosell, Kimmo Kaski, Fabio Babiloni
  University: Univ. Politècnica de Catalunya, Barcelona
  Date: 18 July 2003
  Title: EEG-based Analysis for the Design of Adaptive Brain Interfaces
- **Ph.D. candidate:** Liva Ralaivola
  Committee member: Samy Bengio
  University: Paris 6, Univ. Pierre et Marie Curie
  Date: 17 December 2003
  Title: Modélisation et apprentissage de systèmes et de concepts dynamiques
- **Ph.D. candidate:** Shafi Ullah Khan
  Committee member: Hynek Hermansky
  University: Indian Institute of Technology, Kampur, India
  Title: Articulatory Phonetic Feature Based Neural Network for Continuous ASR
- **Ph.D. candidate:** Chengyi Zheng
  Committee member: Hynek Hermansky
  University: Oregon Health & Sciences University, Portland, Oregon
  Date: 21 November 2003
  Title: Information fusion in ASR

7.4 Courses

- **Title:** Speech and Language Engineering
  Lecturer and Director of the course: Prof. H Bourlard
  School: EPFL, Postgraduate
- **Title:** Decision, estimation and statistical pattern recognition: Application to speech recognition
  Lecturer: Prof. H Bourlard
  School: EPFL, DI/DSC Predoctoral School
- **Title:** Speech Processing
  Lecturer: Prof. H Bourlard
  School: EPFL, Undergraduate (2nd cycle)
- **Title:** Statistical Machine Learning
  Lecturer: Dr Samy Bengio
  School: IDIAP
- **Title:** Signal Processing for Multimedia Engineering
  Lecturer: Prof. H. Hermansky
  School: Graduate School, OGI School of OHSU, Portland, Oregon

7.5 Short term student projects

- **Trainee:** Aïssa Ait-Hassou
  Supervisor: Hervé Bourlard
  University/School: Faculté Polytechnique de Mons (B)
  Date: February 2003 to May 2003
- **Trainee:** Guillermo Zapata Aradilla
  Supervisor: John Dines
  University/School: Catalunya (S)
  Date: September 2003 to February 2004
- **Trainee:** François Bessard
  Supervisor: Sébastien Marcel
  University/School: Univ. de Fribourg (CH)
  Date: July 2003 to December 2003
- **Trainee:** Emmanuel Gabbud
  Supervisor: Hervé Bourlard
  University/School: EPFL (CH)
  Date: March 2003 to May 2003
- **Trainee:** Frédéric Kottelat
  Supervisor: Hervé Bourlard
  University/School: EPFL / Eurecom (F)
  Date: January 2003 to June 2003

- **Trainee:** Jérôme Kowalczyk
  Supervisor: Hervé Bourlard
  University/School: Ecole Nouvelle d'Ingénieurs en Communication (F)
  Date: December 2003 to June 2004
- **Trainee:** Jean-Sébastien Senécal
  Supervisor: Samy Bengio
  University/School: IRO, Montréal (CA)
  Date: May 2003 to September 2003
- **Trainee:** Julien Tiphaigne
  Supervisor: Sébastien Marcel
  University/School: Ecole Nouvelle d'Ingénieurs en Communication (F)
  Date: December 2003 to June 2004

7.6 Other student projects

- **Trainee:** Alain Anthamatten
  Committee member: Pierre Dal Pont, Jean-Albert Ferrez
  University/School: HEVs
  Date: September 2002 - January 2003
  Title: Etude de l'impact économique de l'IDIAP et d'IM2

  Trainee: Bastien Crettol
  Committee member: Frank Formaz
  University/School: ESIS
  Date: September 2003 - January 2004
  Title: gestionMat

8 Scientific Activities

8.1 Editorship

Prof. Hervé Bourlard is

- Member of the Editorial Board, Speech Communication
- Member of the Editorial Board, Intl. Journal of Pattern Recognition and Artificial Intelligence
- Member of the Editorial Board, Journal of Negative Results in Speech and Audio, http://journal.speech.cs.cmu.edu/
- Member of the Editorial Board, Futur(e)

Dr Samy Bengio is

- Associate Editor for the Journal of Computational Statistics

Prof. Hynek Hermansky is

- Membor of the Editorial Board, Speech Communication
- Associate Editor IEEE Transactions on Speech and Audio
- Editor of Special Issue of EURASIP JASP on Anthropomorphic processing of audio and speech

8.2 Scientific and Technical Committees

Prof. Hervé Bourlard is:

- Member of the Board of Trustees, Intl. Computer Science Institute, Berkeley, CA, USA.
- Member of the European Information Society Technology Advisory Group (ISTAG)
- Member of the Foundation Council of the Swiss Network for Innovation
- Fellow of the Engineering and Physical Sciences Research Council (EPSRC), UK
- Chairman of the evaluation panel for European Research and Training Network (RTN) proposals
- Member of the Advisory Council of ISCA (International Speech Communication Association)
- Member of the IEEE Technical Committee on Neural Network Signal Processing
- Member of the Advisory Board of the European Speech Technology Network
- General Chairman, Eurospeech 2003, Geneva
- Program and/or scientific committee member of numerous conferences
- Expert for several European projects

Dr Samy Bengio is

- Program committee member: 4th International Conference on Audio and Video Based Biometric Person Authentication (AVBPA'2003)
- Program Committee member: Neural Information Processing Systems (NIPS'2003)
- Program Committee member: International Conference on Machine Learning (ICML'2004)

- Program Committee member: NNSP'2003, NNSP'2004
- Scientific Committee member: European Symposium on Artificial Neural Networks (ESANN'2004)

Dr Daniel Gatica-Perez is

- Co-organizer, Special Session on Smart Meeting Rooms. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2003.
- Multimedia Processing Area Chair, IEEE Int. Conference on Multimedia and Expo (ICME), 2004.
- Program Committee Member: IEEE Int. Workshop on Neural Networks for Signal Processing (NNSP), 2003. ACM Int. Conf. on Multimedia (ACM MM), 2004. IEEE Int. Workshop on Technology for Education in Developing Countries (TEDC), 2004.

Prof. Hynek Hermansky is/was:

- Member of the Board International Speech Communication Association
- Member of the IEEE Technical Committee on Speech Processing Processing
- Executive Chairman, 2003 International Conference on Text, Speech and Dialogue, Czech Republic
- Member of the Scientific Committee on Speaker Odyssey 2004
- Expert for European Community

Dr José del R. Millán is

- Member of the Programme Committee of the 14th European Conf. on Machine Learning
- Evaluator of the European IST Call on Cognitive Systems.

Dr Jean-Marc Odobez is

- Program Committee Member: IEEE Int. Workshop on Neural Networks for Signal Processing (NNSP), 2003. IAPR Int. Conference Pattern Recognition Workshop : Learning for Adaptable Visual Systems, 2004.

8.3 Short Term Visits

- **Location:** University of Nijmegen

  Visitor: Hynek Hermansky

  Date: 25 Nov 2003

- **Location:** University of Illinois

  Visitor: Hynek Hermansky

  Date: October 2003

- **Location:** France Telecom R&D, Neural Networks for Telecoms Laboratory (DTL/TIC/TNT), Lannion France

  Visitor: Sebastien Marcel

  Date: from 3 nov 2003 to 7 nov 2003

8.4 Scientific Presentations (other than conferences)

In this section, we brielfy list the scientific events and external (e.g., invited) talks, other than conferences, and which did not necessarily result in a publication.

- **Event:** Spring school of the 3rd cycle Romand in statistics

  Date: March 2003

  Speaker: Samy Bengio

  Title: Statistical Machine Learning

- **Event:** University of Illinois

  Date: October 2003

  Speaker: Hynek Hermansky

  Title: Towards Human-Like Engineering

- **Event:** Nijmegen University

  Date: November 2003

  Speaker: Hynek Hermansky

  Title: Vicious Spectral Envelope

- **Event:** Invited seminar, Centre for Advanced Studies, Univ. Santiago de Compostela

  Date: January 9, 2003

  Speaker: José del R. Millán

  Title: Adaptive Brain Interfaces for Communication and Control

- **Event:** Annual Plenary Meeting of the European Medical Research Councils, European Science Foundation, Berlin

  Date: April 24, 2003

  Speaker: José del R. Millán

  Title: Brain-Computer Interfaces

- **Event:** Inst. de Sociologie et Antropologie, Univ. de Lausanne

  Date: January 21, 2003

  Speaker: José del R. Millán

  Title: Merging Brain and Machines

- **Event:** Invited seminar, Biomedical Engineering Research Centre, Technical Univ. of Catalonia (UPC), Barcelona

  Date: July 17, 2003

  Speaker: José del R. Millán

  Title: Non-invasive brain-actuated control of a mobile robot

- **Event:** Vision Book project, European Commission's Information Society DG, Brussels

  Date: December 3, 2003

  Speaker: José del R. Millán

  Title: Tapping the mind or resonating minds?

9 Publications (2002 and 2003)

9.1 Books and Book Chapters

[1] H. BOURLARD, T. ADALI, S. BENGIO, J. LARSEN, AND S. DOUGLAS, eds., *Proceedings of the Twelfth IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, IEEE Press, 2002.

[2] H. BOURLARD AND S. BENGIO, *Hidden markov models and other finite state automata for sequence processing*, in The Handbook of Brain Theory and Neural Networks: The Second Edition, M. A. Arbib, ed., The MIT Press, 2002.

[3] H. BOURLARD, S. BENGIO, AND K. WEBER, *Towards robust and adaptive speech recognition models*, in Mathematical Foundations of Speech Processing and Recognition, M. Ostendorf, S. Khudanpur, and R. Rosenfeld, eds., Institute for Mathematics and its Applications (IMA) Series, Springer-Verlag, 2002.

[4] J. MILLÁN, *Brain-computer interfaces*, in The Handbook of Brain Theory and Neural Networks: The Second Edition, M. A. Arbib, ed., The MIT Press, 2002.

[5] J. MILLÁN, *Robot navigation*, in The Handbook of Brain Theory and Neural Networks: The Second Edition, M. A. Arbib, ed., The MIT Press, 2002.

9.2 Articles in International Journals

[1] J. AJMERA, I. MCCOWAN, AND H. BOURLARD, *Robust speaker change detection*, IEEE Signal Processing Letters (to appear), (2003).

[2] J. AJMERA, I. MCCOWAN, AND H. BOURLARD, *Speech/music discrimination using entropy and dynamism features in a hmm classification framework*, Speech Communication, 40 (2003), pp. 351–363.

[3] S. BENGIO, *Multimodal speech processing using asynchronous hidden markov models*, Information Fusion, 5 (2004).

[4] S. BENGIO, C. MARCEL, S. MARCEL, AND J. MARIÉTHOZ, *Confidence measures for multimodal identity verification*, Information Fusion, 3 (2002), pp. 267–276.

[5] F. CAMASTRA AND A. VINCIARELLI, *Estimating the intrinsic dimension of data with a fractal-based method*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), pp. 1404–1407.

[6] F. CAMASTRA AND A. VINCIARELLI, *Combining Neural Gas and Learning Vector Quantization for cursive character recognition*, Neurocomputing, 51 (2003), pp. 147–159.

[7] D. CHEN, J.-M. ODOBEZ, AND H. BOURLARD, *Text Detection and Recognition in Images and Videos*, Pattern Recognition, 37 (2004), pp. 595–609.

[8] D. CHEN, J.-M. ODOBEZ, AND J.-P. THIRAN, *A localization/verification scheme for finding text in images and video frames based on contrast independent features and machine learning methods*, to appear in Signal Processing: Image Communication, (2003).

[9] R. COLLOBERT, S. BENGIO, AND Y. BENGIO, *A parallel mixture of SVMs for very large scale problems*, Neural Computation, 14 (2002), pp. 1105–1114.

[10] CONRAD SANDERSON AND KULDIP K. PALIWAL, *Fast features for face authentication under illumination direction changes*, Pattern Recognition Letters, 24 (2003), pp. 2409–2419.

[11] CONRAD SANDERSON AND KULDIP K. PALIWAL, *Structurally noise resistant classifier for multi-modal person verification*, Pattern Recognition Letters, 24 (2003), pp. 3089–3099.

[12] B. FASEL AND J. LUETTIN, *Automatic Facial Expression Analysis: A Survey*, Pattern Recognition, 36 (2003), pp. 259–275.

[13] D. GATICA-PEREZ, A. LOUI, AND M.-T. SUN, *Finding structure in home videos by probabilistic hierarchical clustering*, IEEE Transactions on Circuits and Systems for Video Technology, 13 (2003), pp. 539–548.

[14] I. LAPIDOT AND H. GUTERMAN, *Dichotomy between clustering performance and minimum distortion in piecewise-dependent-data (PDD) clustering*, to be published in IEEE Signal Processing Letters, (2003).

[15] I. MCCOWAN AND H. BOURLARD, *Microphone array post-filter based on noise field coherence*, IEEE Transactions on Speech and Audio Processing, 11 (2003).

[16] J. MILLÁN, *Adaptive brain interfaces*, Communications of the ACM, 46 (2003).

[17] J. MILLÁN AND J. M. NO, *Asynchronous BCI and local neural classifiers: An overview of the Adaptive Brain Interface project*, to appear in IEEE Trans. on Neural Systems and Rehabilitation Engineering, Special Issue on Brain-Computer Interface Technology, (2003).

[18] S. MOELLER AND H. BOURLARD, *Analytic assessment of telephone transmission impact on asr performance using a simulation model*, Speech Communication, (2002).

[19] C. N. A. H. PERE PUJOL, SUSAGNA POL AND H. BOURLARD, *Comparison and Combination of Features in a Hybrid hmm/mlp and a hmm/gmm Speech Recognition System*, to be published in IEEE Transactions on Speech and Audio Processing, (2003).

[20] T. A. STEPHENSON, M. MAGIMAI-DOSS, AND H. BOURLARD, *Speech recognition with auxiliary information*, to be published in IEEE Trans. on Speech and Audio Processing, (2003).

[21] A. VINCIARELLI, *A survey on off-line cursive word recognition*, Pattern Recognition, 35 (2002), pp. 1433–1446.

[22] A. VINCIARELLI AND S. BENGIO, *Writer adaptation techniques in HMM based off-line cursive script recognition*, Pattern Recognition Letters, 23 (2002), pp. 905–916.

[23] S. A.VINCIARELLI AND H.BUNKE, *Offline recognition of unconstrained handwritten texts using HMMs and statistical language models*, accepted for publication by IEEE Transactions on Pattern Analysis and Machine Intelligence, (2003).

[24] K. WEBER, S. IKBAL, S. BENGIO, AND H. BOURLARD, *Robust Speech Recognition and Feature Extraction Using HMM2*, Computer Speech & Language, 17 (2003), pp. 195–211.

9.3 Articles in Conference Proceedings

[1] J. AJMERA, H. BOURLARD, I. LAPIDOT, AND I. MCCOWAN, *Unknown-multiple speaker clustering using hmm*, in ICSLP, Denver, Colorado, 2002, pp. 573–576.

[2] J. AJMERA, G. LATHOUD, AND I. MCCOWAN, *Clustering and segmenting speakers and their locations in meetings*, in ICASSP, 2004.

[3] J. AJMERA, I. MCCOWAN, AND H. BOURLARD, *Robust hmm-based speech/music segmentation*, in ICASSP, Orlando, Florida, 2002, pp. 1746–1749.

[4] J. AJMERA AND C. WOOTERS, *A robust speaker clustering algorithm*, in IEEE Automatic Speech Recognition Understanding Workshop.

[5] E. BAILLY-BAILLIÈRE, S. BENGIO, F. BIMBOT, M. HAMOUZ, J. KITTLER, J. MARIÉTHOZ, J. MATAS, K. MESSER, V. POPOVICI, F. PORÉE, B. RUIZ, AND J.-P. THIRAN, *The BANCA database and evaluation protocol*, in 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag, 2003.

[6] M. BARNARD, J.-M. ODOBEZ, AND S. BENGIO, *Multi-modal audio-visual event recognition for football analysis*, in Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP), Toulouse, France, September 2003.

[7] S. BENGIO, *An asynchronous hidden markov model for audio-visual speech recognition*, in Advances in Neural Information Processing Systems, NIPS 15, S. Becker, S. Thrun, and K. Obermayer, eds., Vancouver, Canada, 2003, MIT Press.

[8] S. BENGIO, *Multimodal authentication using asynchronous HMMs*, in 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag, 2003.

[9] M. F. BENZEGHIBA AND H. BOURLARD, *User-Customized Password HMM Based Speaker Verification*, in Proceedings of the COST275 Workshop on the Advent of Biometrics on the Internet, Rome, Italy, 2002, pp. 103–106.

[10] M. F. BENZEGHIBA AND H. BOURLARD, *User-Customized Password Speaker Verification based on HMM/ANN and GMM Models*, in International Conference on Spoken Language Processing (ICSLP 2002), Denver, CO, USA, 2002, pp. 1325–1328.

[11] M. F. BENZEGHIBA AND H. BOURLARD, *Confidence Measures in Multiple pronunciations Modeling For Speaker Verification*, IDIAP-RR 53, IDIAP, 2003.

[12] M. F. BENZEGHIBA AND H. BOURLARD, *Hybrid HMM/ANN and GMM Combination for User-Customized Password Speaker Verification*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), 2003.

[13] M. F. BENZEGHIBA AND H. BOURLARD, *On the Combination of Speech and Speaker Recognition*, in European Conference On Speech, Communication and Technology (EUROSPEECH'03), Geneva, Switzerland, 2003, pp. 1361–1364.

[14] H. BOURLARD, I. MCCOWAN, D. GATICA-PEREZ, S. BENGIO, AND D. MOORE(2004), *Modeling Human Interactions in Meetings based on Audio and Visual Information*, invited talk, Proc. of Special Workshop in Maui (SWIM), Lectures by Masters in Speech Processing, Maui, Hawaii, January 12-14. 2004.

[15] F. CARDINAUX AND S. MARCEL, *Face verification using MLP and SVM*, in XI Journees NeuroSciences et Sciences pour l'Ingenieur (NSI 2002), no. 21, La Londe Les Maures, France, 15-19 September 2002.

[16] F. CARDINAUX, C. SANDERSON, AND S. MARCEL, *Comparison of MLP and GMM classifiers for face verification on XM2VTS*, in 4th International Conference on Audio- and Video-Based Biometric Person Authentication, no. 10, University of Surrey, Guildford, UK, June 9-11 2003.

[17] D. CHEN AND J.-M. ODOBEZ, *Sequential Monte Carlo Video Text Segmentation*, in ICIP, Sep. 2003.

[18] D. CHEN, J.-M. ODOBEZ, AND H. BOURLARD, *Text Segmentation and Recognition in Complex Background Based on Markov Random Field*, in Int. Conf. Pattern Recognition 2002, Quebec city, Canada, Oct 2002.

[19] F. CINCOTTI, A. SCIPIONE, A. TINIPERI, M. M. D. MATTIA, J. MILLÁN, S. SALINARI, L. BIANCHI, AND F. BABILONI, *Comparison of different feature classifiers for brain computer interfaces*, in Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering, Capri, Italy, March 2003.

[20] R. COLLOBERT AND S. BENGIO, *A gentle hessian for efficient gradient descent*, in IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP, 2004.

[21] R. COLLOBERT, S. BENGIO, AND Y. BENGIO, *A parallel mixture of SVMs for very large scale problems*, in Advances in Neural Information Processing Systems, NIPS 14, T. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, 2002.

[22] R. COLLOBERT, Y. BENGIO, AND S. BENGIO, *Scaling large learning problems with hard parallel mixtures*, in International Workshop on Pattern Recognition with Support Vector Machines, SVM'2002, 2002.

[23] CONRAD SANDERSON AND KULDIP K. PALIWAL, *Noise Resistant Audio-Visual Verification via Structural Constraints*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003.

[24] CONRAD SANDERSON AND SAMY BENGIO, *Augmenting Frontal Face Models for Non-Frontal Verification*, in Proceedings of the 2003 Workshop on Multimodal User Authentication (MMUA'03), Santa Barbara, California, December 2003.

[25] CONRAD SANDERSON AND SAMY BENGIO, *Robust Features for Frontal Face Authentication in Difficult Image Conditions*, in Proceedings of 4th International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA-03), University of Surrey, Guildford, United Kingdom, June 2003.

[26] J. CZYZ, S. BENGIO, C. MARCEL, AND L. VANDENDORPE, *Scalability analysis of audio-visual person identity verification*, in 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag, 2003.

[27] I. M. DANIEL GATICA-PEREZ, GUILLAUME LATHOUD AND J.-M. ODOBEZ, *A mixed-state i-particle filter for multi-camera speaker tracking*, in IEEE Int. Conf. on Computer Vision Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC), 2003.

[28] C. DIMITRAKAKIS AND S. BENGIO, *Boosting hmms with an application to speech recognition*, in IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP, 2004.

[29] C. DIMITRAKAKIS AND S. BENGIO, *Online policy adaptation for ensemble classifiers*, in 12th European Symposium on Artificial Neural Networks, ESANN 04, 2004.

[30] FABIEN CARDINAUX, CONRAD SANDERSON, AND SAMY BENGIO, *Face Verification Using Adapted Generative Models*, in The 6th International Conference on Automatic Face and Gesture Recognition, FG2004, Seoul, Korea, 2004, IEEE.

[31] B. FASEL, *Facial Expression Analysis using Shape and Motion Information Extracted by Convolutional Neural Networks*, in International IEEE Workshop on Neural Networks for Signal Processing (NNSP 02), Martigny, Switzerland, sep 2002, pp. 607–616.

[32] B. FASEL, *Head-Pose Invariant Facial Expression Recognition using Convolutional Neural Networks*, in International IEEE Conference on Multimodal Interfaces (ICMI 02), Pittsburgh, USA, oct 2002, pp. 529–534.

[33] B. FASEL, *Mutliscale Facial Expression Recognition using Convolutional Neural Networks*, in Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 02), Ahmedabad, India, dec 2002.

[34] B. FASEL, *Robust Face Analysis using Convolutional Neural Networks*, in Proceedings of the International Conference on Pattern Recognition (ICPR 02), vol. 2, Quebec, Canada, aug 2002, pp. 40–43.

[35] D. GATICA-PEREZ, G. LATHOUD, I. MCCOWAN, J.-M. ODOBEZ, AND D. MOORE, *Audio-visual speaker tracking with importance particle filters*, in IEEE International Conference on Image Processing (ICIP), 2003.

[36] D. GATICA-PEREZ, I. MCCOWAN, M. BARNARD, S. BENGIO, AND H. BOURLARD, *On automatic annotation of meeting databases*, in IEEE International Conference on Image Processing (ICIP), 2003.

[37] D. GATICA-PEREZ AND M.-T. SUN, *Linking objects in videos by importance sampling*, in IEEE International Conference on Multimedia and Expo, 2002.

[38] D. GATICA-PEREZ AND M.-T. SUN, *Object localization in metric spaces for video linking*, in IEEE Workshop on Motion and Video Computing, 2002.

[39] D. GATICA-PEREZ, M.-T. SUN, AND A. LOUI, *Probabilistic home video structuring: Feature selection and performance evaluation*, in IEEE International Conference on Image Processing, 2002.

[40] N. GILARDI, S. BENGIO, AND M. KANEVSKI, *Conditional gaussian mixture models for environmental risk mapping*, in IEEE International Workshop on Neural Networks for Signal Processing (NNSP), 2002.

[41] R. GRAVE DE PERALTA, S. GONZÁLEZ, J. MILLÁN, T. PUN, AND C. MICHEL, *Direct non-invasive brain computer interfaces*, in Proceedings of the 9th International Conference on Functional Mapping of the Human Brain, New York, USA, June 2003.

[42] M. GUILLEMOT, P. WELLNER, D. GATICA-PEREZ, AND J-M.ODOBEZ, *A hierarchical keyframe user interface for browsing video over the internet*, in Proceedings of the 9th International Conference on Human-Computer Interaction (INTERACT-2003), R. M., W. J., and M. M., eds., Zurich, Switzerland, 2003, IOS Press.

[43] E. GYSELS, J. MILLÁN, S. CHIAPPA, AND P. CELKA, *Studying phase synchrony for classification of mental tasks in brain machine interfaces*, in Proceedings of the Conference of the International Society for Brain ElectromagneticTopography, Santa Fe, USA, November 2003.

[44] H. HERMANSKY, *TRAP-TANDEM: Data-driven extraction of temporal features from speech*, in large part published in Proceedings of ASRU-2003, no. 50, Martigny, Switzerland, 2003.

[45] H. HERMANSKY AND H. BOURLARD, *Some Emerging Concepts in Speech Recognition*, invited talk, Proc. of Special Workshop in Maui (SWIM), Lectures by Masters in Speech Processing, Maui, Hawaii, January 12-14. 2004.

[46] S. IKBAL, H. HERMANSKY, AND H. BOURLARD, *Nonlinear Spectral Transformations for Robust Speech Recognition*, in Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop 2003, St. Thomas, U.S. Virgin Islands, USA, Dec 2003.

[47] S. IKBAL, H. MISRA, AND H. BOURLARD, *Phase AutoCorrelation (PAC) derived Robust Speech Features*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003.

[48] S. IKBAL, H. MISRA, H. BOURLARD, AND H. HERMANSKY, *Phase AutoCorrelation (PAC) features in Entropy based Multi-Stream for Robust Speech Recognition*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-04), Montreal, Canada, May 2004.

[49] S. IKBAL, K. WEBER, AND H. BOURLARD, *Speaker Normalization using HMM2*, in Proceedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP-02), Martigny, Switzerland, September 2002, pp. 647–656.

[50] G. LATHOUD AND I. MCCOWAN, *Location Based Speaker Segmentation*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003.

[51] G. LATHOUD, I. A. MCCOWAN, AND D. C. MOORE, *Segmenting Multiple Concurrent Speakers Using Microphone Arrays*, in Proceedings of Eurospeech 2003, Geneva, Switzerland, September 2003.

[52] Q. LE AND S. BENGIO, *Client dependent gmm-svm models for speaker verification*, in International Conference on Artificial Neural Networks, ICANN/ICONIP 2003, Springer Verlag, 2003.

[53] M. MAGIMAI.-DOSS, S. BENGIO, AND H. BOURLARD, *Joint decoding for phoneme-grapheme continuous speech recognition*, in Proceedings of ICASSP, Montreal, Canada, 2004.

[54] M. MAGIMAI.-DOSS, T. A. STEPHENSON, AND H. BOURLARD, *Using pitch frequency information in speech recognition*, in Proceedings of Eurospeech, vol. 4, Geneva, Switzerland, September 2003, pp. 2525–2528.

[55] M. MAGIMAI.-DOSS, T. A. STEPHENSON, H. BOURLARD, AND S. BENGIO, *Phoneme-grapheme based speech recognition system*, in Proceedings of IEEE ASRU, U.S. Virgin Islands, USA, 2003.

[56] S. MARCEL AND S. BENGIO, *Improving face verification using skin color information*, in Proceedings of the 16th International Conference on Pattern Recognition, IEEE Computer Society Press, 2002.

[57] S. MARCEL, C. MARCEL, AND S. BENGIO, *A state-of-the-art Neural Network for robust face verification*, in Proceedings of the COST275 Workshop on The Advent of Biometrics on the Internet, Rome, Italy, 2002.

[58] MARIÉTHOZ, J. AND BENGIO, S., *A comparative study of adaptation methods for speaker verification*, in International Conference on Spoken Language Processing ICSLP, Denver, CO, USA, September 2002, pp. 581–584.

[59] I. MCCOWAN, S. BENGIO, D. GATICA-PEREZ, G. LATHOUD, F. MONAY, D. MOORE, P. WELLNER, AND H. BOURLARD, *Modeling human interaction in meetings*, in Proceedings of International Conference on Acoustics, Speech and Signal Processing, Hong Kong, April 2003.

[60] I. MCCOWAN, A. MORRIS, AND H. BOURLARD, *Robust speech recognition with small microphone arrays using the missing data approach*, IDIAP-RR 09, IDIAP, Martigny, Switzerland, 2002.

[61] J. MILLÁN, *Adaptive brain interfaces for communication and control*, in Proceedings of the 10th International Conference on Human-Computer Interaction, Crete, Greece, June 2003.

[62] J. MILLÁN, *Restoring locomotion with a thought controlled mobile robot*, in Proceedings of the 4th Forum of European Neuroscience, Lisbon, Portugal, June 2004.

[63] J. MILLÁN, F. RENKENS, J. MOURIÑO, AND W. GERSTNER, *Non-invasive brain-actuated control of a mobile robot*, in Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 2003.

[64] H. MISRA, H. BOURLARD, AND V. TYAGI, *New entropy based combination rules in HMM/ANN multistream ASR*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, April 2003.

[65] H. MISRA AND A. C. MORRIS, *Confusion Matrix Based Entropy Correction in Multi-stream Combination*, in Proceedings of Eurospeech, Geneve, Switzerland, September 2003.

[66] F. MONAY AND D. GATICA-PEREZ, *On image auto-annotation with latent space models*, in Proc. ACM Int. Conf. on Multimedia (ACM MM), Nov. 2003.

[67] D. MOORE AND I. MCCOWAN, *Microphone array speech recognition : Experiments on overlapping speech in meetings*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003.

[68] A. MORRIS, S. PAYNE, AND H. BOURLARD, *Low cost duration modelling for noise robust speech recognition*, in Proc. ICSLP, Denver, Colorado, USA, September 16-20 2002.

[69] J. MOURIÑO, S. CHIAPPA, R. JANÉ, AND J. MILLÁN, *Evolution of the mental states operating a brain-computer interface*, in Proceedings of the International Federation for Medical and Biological Engineering, Vienna, Austria, December 2002.

[70] J.-M. ODOBEZ AND S. BA, *Modélisation implicite du mouvement en suivi par filtrage de monte carlo séquentiel*, in GRETSI conference, Signal and Image Processing,, Paris, France, 2003.

[71] J.-M. ODOBEZ, S. BA, AND D. GATICA-PEREZ, *An Implicit Motion Likelihood for Tracking with Particle Filters*, in British Machine Vision Conference (BMVC), Lecture Notes in Computer Science, Norwich, UK, 2003, Springer Verlag.

[72] J.-M. ODOBEZ AND D. CHEN, *Video Text Recognition based on Markov Random Field and Grayscale Consistency Constraint*, in Int. Conf. Image Processing 2002, Rochester, NY, USA, Sept 2002.

[73] J.-M. ODOBEZ, D. GATICA-PEREZ, AND M. GUILLEMOT, *Spectral Structuring of Home Videos*, in International Conference on Image and Video Retrieval (CIVR'03), Lecture Notes in Computer Science, Urbana-Champaign, USA, July 2003, Springer Verlag.

[74] J.-M. ODOBEZ, D. GATICA-PEREZ, AND M. GUILLEMOT, *Video Shot Clustering using Spectral Methods*, in 3rd Workshop on Content-Based Multimedia Indexing (CBMI), Rennes, France, 2003.

[75] J. B. PEDRO QUELHAS, *Vessel segmentation and branching detection using an adaptive profile kalman filter in retinal blood vessel structure analysis*, in Pattern Recognition and Image Analysis: First Iberian Conference, IbPRIA 2003, Springer-Verlag LNCS, vol. 2652, June 2003.

[76] N. POH AND S. BENGIO, *Non-linear variance reduction techniques in biometric authentication*, IDIAP-RR 26, IDIAP, 2003.

[77] N. POH, S. BENGIO, AND J. KORCZAK, *A multi-sample multi-source model for biometric authentication*, in IEEE International Workshop on Neural Networks for Signal Processing (NNSP), 2002.

[78] N. POH, S. MARCEL, AND S. BENGIO, *Improving face authetication using virtual samples*, in IEEE International Conference on Acoustics, Speech, and Signal Processing, no. 40, 2003.

[79] C. SANDERSON, S. BENGIO, H. BOURLARD, J. MARIETHOZ, R. COLLOBERT, M. F. BENZEGHIBA, F. CARDINAUX, AND S. MARCEL, *Speech & Face Based Biometric Authentication at IDIAP*, in Proceedings of the 2003 IEEE International Conference on Multimedia & Expo (ICME-03), Baltimore, Maryland, July 2003.

[80] S. SIVADAS AND H. HERMANSKY, *On Use of Task Independent Training Data in Tandem Feature Extraction*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-04), Montreal, Canada, May 2004.

[81] T. A. STEPHENSON, J. ESCOFET, M. MAGIMAI-DOSS, AND H. BOURLARD, *Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables*, in 2002 IEEE International Workshop on Neural Networks for for Signal Processing (NNSP 2002), Martigny, Switzerland, September 2002, pp. 637–646.

[82] T. A. STEPHENSON, M. MAGIMAI-DOSS, AND H. BOURLARD, *Auxiliary variables in conditional Gaussian mixtures for automatic speech recognition*, in Seventh International Conference on Spoken Language Processing (ICSLP 2002), vol. 4, Denver, CO, USA, September 2002, pp. 2665–2668.

[83] T. A. STEPHENSON, M. MAGIMAI-DOSS, AND H. BOURLARD, *Mixed Bayesian networks with auxiliary variables for automatic speech recognition*, in International Conference on Pattern Recognition (ICPR 2002), vol. 4, Quebec City, PQ, Canada, August 2002, pp. 293–296.

[84] T. A. STEPHENSON, M. MAGIMAI-DOSS, AND H. BOURLARD, *Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks*, in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), vol. I, Hong Kong, April 2003, pp. 20–23.

[85] V. TYAGI, I. MCCOWAN, H. BOURLARD, AND H. MISRA, *Mel-cepstrum modulation spectrum (mcms) features for robust asr*, in IEEE ASRU, 2003.

[86] V. TYAGI, I. MCCOWAN, H. BOURLARD, AND H. MISRA, *On factorizing spectral dynamics for robust speech recognition*, in Eurospeech, 2003.

[87] A. VINCIARELLI AND S. BENGIO, *Offline cursive word recognition using continuous density Hidden Markov Models trained with PCA or ICA features*, in Proceedings of International Conference on Pattern Recognition, vol. III, Quebec City (Canada), 2002, pp. 81–84.

[88] A. VINCIARELLI AND S. BENGIO, *Writer adaptation techniques in HMM based off-line cursive script recognition*, in Proceedings of 8^{th} International Conference on Frontiers on Handwriting Recognition, Niagara on the Lake (Canada), 2002, pp. 287–291.

[89] A. VINCIARELLI, S. BENGIO, AND H. BUNKE, *Offline recognition of large vocabulary cursive handwritten text*, in Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 2003, pp. 1101–1104.

[90] K. WEBER, S. BENGIO, AND H. BOURLARD, *Increasing Speech Recognition Noise Robustness with HMM2*, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 02), Orlando, Florida, USA, May 2002, pp. I.929–932.

[91] K. WEBER, F. DE WET, B. CRANEN, L. BOVES, S. BENGIO, AND H. BOURLARD, *Evaluation of Formant-Like Features for ASR*, in International Conference on Spoken Language Processing (ICSLP 2002), Denver, CO, USA, September 2002, pp. 2101–2104.

9.4 IDIAP Research Reports (submitted for publication or not published)

[1] J. AJMERA, H. BOURLARD, AND I. LAPIDOT, *Improved unknown-multiple speaker clustering using hmm*, IDIAP-RR 23, IDIAP, Martigny, Switzerland, 2002.

[2] A.POZDNOUKHOV, *The analysis of kernel ridge regression learning algorithm.*, IDIAP-RR 54, IDIAP, Martigny, Switzerland, 2002.

[3] S. BA AND J.-M. ODOBEZ, *A Probabilistic Framework for Joint Head Tracking and Pose Estimation*, IDIAP-RR 78, IDIAP, Martigny, Switzerland, 2003.

[4] M. BARNARD AND J.-M. ODOBEZ, *Robust playfield segmentation using map adaptation*, IDIAP-RR 77, IDIAP, 2003.

[5] S. BENGIO, F. BIMBOT, J. MARIÉTHOZ, V. POPOVICI, F. PORÉE, E. BAILLY-BAILLIÈRE, G. MATAS, AND B. RUIZ, *Experimental protocol on the BANCA database*, IDIAP-RR 05, IDIAP, 2002.

[6] S. BENGIO, M. KELLER, AND J. MARIÉTHOZ, *The expected performance curve*, IDIAP-RR 85, IDIAP, 2003.

[7] S. BENGIO AND J. MARIÉTHOZ, *The expected performance curve: a new assessment measure for person authentication*, IDIAP-RR 84, IDIAP, 2003.

[8] S. BENGIO AND J. MARIÉTHOZ, *A statistical significance test for person authentication*, IDIAP-RR 83, IDIAP, 2003.

[9] Y. BENGIO AND J.-S. SENÉCAL, *Adaptive importance sampling to accelerate training of a neural probabilistic language model*, IDIAP-RR 35, IDIAP, 2003.

[10] F. CAMASTRA AND A. VINCIARELLI, *Estimating the intrinsic dimension of data with a fractal-based method*, IDIAP-RR 02, IDIAP, 2002.

[11] F. CARDINAUX AND S. MARCEL, *Face verification using MLP and SVM*, IDIAP-RR 21, IDIAP, 2002.

[12] F. CARDINAUX, C. SANDERSON, AND S. MARCEL, *Comparison of MLP and GMM classifiers for face verification on XM2VTS*, IDIAP-RR 10, IDIAP, 2003.

[13] D. CHEN AND J.-M. ODOBEZ, *A New Method of Contrast Normalization for Verification of Extracted Video Text having Complex Backgrounds*, IDIAP-RR-02 16, IDIAP, Apr 2002.

[14] D. CHEN AND J.-M. ODOBEZ, *Comparison of Support Vector Machine and Neural Network for Text Texture Verification*, IDIAP-RR-02 19, IDIAP, Martigny, Apr 2002.

[15] D. CHEN AND J.-M. ODOBEZ, *Video text segmentation using particle filters*, IDIAP-RR-03 43, IDIAP, May 2003.

[16] S. CHIAPPA AND S. BENGIO, *Nonlinear analysis of cognitive and motor-related eeg signals*, IDIAP-RR 14, IDIAP, 2003.

[17] R. COLLOBERT AND S. BENGIO, *A new margin-based criterion for efficient gradient descent*, IDIAP-RR 16, IDIAP, 2003.

[18] R. COLLOBERT AND S. BENGIO, *Links between perceptrons, mlps and svms*, IDIAP-RR 06, IDIAP, 2004.

[19] R. COLLOBERT, S. BENGIO, AND J. MARIÉTHOZ, *Torch: a modular machine learning software library*, IDIAP-RR 46, IDIAP, 2002.

[20] CONRAD SANDERSON, *Face processing & frontal face verification*, IDIAP-RR 20, IDIAP, April 2003.

[21] CONRAD SANDERSON AND SAMY BENGIO, *Face Verification Using Synthesized Non-Frontal Models*, IDIAP-RR 60, IDIAP, November 2003.

[22] CONRAD SANDERSON AND SAMY BENGIO, *Robust Features for Frontal Face Authentication in Difficult Image Conditions*, IDIAP-RR 05, IDIAP, January 2003.

[23] CONRAD SANDERSON AND SAMY BENGIO, *Statistical Transformation Techniques for Face Verification Using Faces Rotated in Depth*, IDIAP-RR 04, IDIAP, February 2004.

[24] F. DE WET, K. WEBER, L. BOVES, B. CRANEN, S. BENGIO, AND H. BOURLARD, *Evaluation of formant-like features for automatic speech recognition*, IDIAP-RR 08, IDIAP, 2003.

[25] C. DIMITRAKAKIS AND S. BENGIO, *Online policy adaptation for ensemble algorithms*, IDIAP-RR 28, IDIAP, 2002.

[26] C. DIMITRAKAKIS AND S. BENGIO, *Online policy adaptation for ensemble classifiers*, IDIAP-RR 69, IDIAP, 2003.

[27] J. ESCOFET CARMONA AND T. A. STEPHENSON, *Automatic speech recognition using dynamic Bayesian networks with the energy as an auxiliary variable*, IDIAP-RR 18, IDIAP, 2003.

[28] N. GILARDI, S. BENGIO, AND M. KANEVSKI, *Estimation of conditional distributions using gaussian mixture models*, IDIAP-RR 03, IDIAP, 2002.

[29] D. GRANGIER AND A. VINCIARELLI, *Making retrieval faster through document clustering*, IDIAP-RR 02, IDIAP, 2004.

[30] A. HAGEN AND A. C. MORRIS, *Recent advances in the multi-stream hmm/ann hybrid approach to noise robust asr*, IDIAP-RR 57, IDIAP, 2002.

[31] JOHNNY MARIÉTHOZ AND SAMY BENGIO, *An alternative to silence removal for text-independent speaker verification*, IDIAP-RR 51, IDIAP, 2003.

[32] A. JUST, O. BERNIER, AND S. MARCEL, *Recognition of isolated complex mono- and bi-manual 3d hand gestures*, IDIAP-RR 63, IDIAP, 2003.

[33] A. JUST, S. MARCEL, O. BERNIER, AND J. VIALLET, *Reconnaissance de gestes 3d bi-manuels*, IDIAP-RR 79, IDIAP, 2003.

[34] M. KELLER AND S. BENGIO, *Theme Topic Mixture Model: A Graphical Model for Document Representation*, IDIAP-RR 05, IDIAP, 2004.

[35] I. LAPIDOT, *Self-organizing-maps with BIC for speaker clustering*, IDIAP-RR 60, IDIAP, Martigny, Switzerland, 2002.

[36] I. LAPIDOT, *What is better: GMM of two gaussians or two clusters with one gaussian?*, IDIAP-RR 56, IDIAP, Martigny, Switzerland, 2002.

[37] I. LAPIDOT AND A. MORRIS, *Extended BIC criterion for model selection*, IDIAP-RR 42, IDIAP, Martigny, Switzerland, 2002.

[38] Q. LE AND S. BENGIO, *Hybrid generative-discriminative models for speech and speaker recognition*, IDIAP-RR 06, IDIAP, 2002.

[39] Q. LE AND S. BENGIO, *Noise robust discriminative models*, IDIAP-RR 40, IDIAP, 2003.

[40] V. LEMAIRE, *Bagging using the vmse cost function*, IDIAP-RR 27, France Telecom Research and Development, 2002.

[41] V. LEMAIRE AND F. CLÉROT, *Som-based clustering for on-line fraud behavior classification: a case study*, IDIAP-RR 30, France Telecom Research and Development, 2002.

[42] M. MAGIMAI.-DOSS, S. BENGIO, AND H. BOURLARD, *Joint decoding for phoneme-grapheme continuous speech recognition*, IDIAP-RR 52, IDIAP, 2003.

[43] M. MAGIMAI.-DOSS, T. A. STEPHENSON, AND H. BOURLARD, *Modelling auxiliary information (pitch frequency) in hybrid HMM/ANN based ASR systems*, IDIAP-RR 62, IDIAP, 2002.

[44] S. MARCEL, *Evaluation protocols and comparative results for the Triesch hand posture database*, IDIAP-RR 50, IDIAP, 2002.

[45] S. MARCEL, *Gestures for multi-modal interfaces: A review*, IDIAP-RR 34, IDIAP, 2002.

[46] S. MARCEL, *Robust face verification using skin color and Neural Networks*, IDIAP-RR 49, IDIAP, 2002.

[47] S. MARCEL, *Face verification using lda and mlp on the banca database*, IDIAP-RR 66, IDIAP, 2003.

[48] S. MARCEL, *Improving face verification using symmetric transformation*, IDIAP-RR 68, IDIAP, 2003.

[49] S. MARCEL, *A symmetric transformation for lda-based face verification*, IDIAP-RR 67, IDIAP, 2003.

[50] I. MCCOWAN, J. AJMERA, AND D. MORRE, *An online system for automatic annotation of audio documents*, IDIAP-RR 39, IDIAP, 2003.

[51] I. MCCOWAN AND H. BOURLARD, *Microphone array post-filter for diffuse noise field*, IDIAP-RR 39, IDIAP, Martigny, Switzerland, 2001.

[52] I. MCCOWAN, D. GATICA-PEREZ, S. BENGIO, AND H. BOURLARD, *Towards computer understanding of human interactions*, IDIAP-RR 45, IDIAP, Martigny, Switzerland, 2003.

[53] J. MILLÁN, *On the need for on-line learning in brain-computer interfaces*, IDIAP-RR 30, IDIAP, Martigny, Switzerland, 2003.

[54] A. C. MORRIS, *Noise pdf transformation in secondary feature processing*, IDIAP-RR 29, IDIAP, 2002.

[55] J.-M. ODOBEZ, S. BA, AND D. GATICA-PEREZ, *An Implicit Motion Likelihood for Tracking with Particle Filters*, IDIAP-RR 15, IDIAP, Martigny, Switzerland, 2003.

[56] J.-M. ODOBEZ, D. GATICA-PEREZ, AND S. BA, *Embedding motion in model-based stochastic tracking*, IDIAP-RR 72, IDIAP, Martigny, Switzerland, 2003.

[57] N. POH AND S. BENGIO, *Variance reduction techniques in biometric authentication*, IDIAP-RR 17, IDIAP, 2003.

[58] N. POH AND S. BENGIO, *Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?*, IDIAP-RR 59, IDIAP, 2003.

[59] N. POH AND S. BENGIO, *Noise-robust multi-stream fusion for text-independent speaker authentication*, IDIAP-RR 01, IDIAP, 2004.

[60] N. POH, C. SANDERSON, AND S. BENGIO, *An investigation of spectral subband centroids for speaker authentication*, IDIAP-RR 62, IDIAP, 2003.

[61] F. PORÉE, J. MARIÉTHOZ, S. BENGIO, AND F. BIMBOT, *The BANCA database and experimental protocol for speaker verification*, IDIAP-RR 13, IDIAP, 2002.

[62] A. POZDNOUKHOV AND S. BENGIO, *From samples to objects in kernel methods*, IDIAP-RR 29, IDIAP, Martigny, Switzerland, 2003.

[63] A. POZDNOUKHOV AND S. BENGIO, *Tangent vector kernels for invariant image classification with svms*, IDIAP-RR 75, IDIAP, Martigny, Switzerland, 2003.

[64] P. QUELHAS AND J.-M. ODOBEZ, *A color and gradient local descriptor fusion scheme for object recognition*, IDIAP-RR 71, IDIAP, 2003.

[65] Y. RODRIGUEZ AND S. MARCEL, *Boosting pixel-based classifiers for face verification*, IDIAP-RR 65, IDIAP, 2003.

[66] C. SANDERSON, S. BENGIO, H. BOURLARD, J. MARIETHOZ, R. COLLOBERT, M. F. BENZEGHIBA, F. CARDINAUX, AND S. MARCEL, *Speech & Face Based Biometric Authentication at IDIAP*, IDIAP-RR 13, IDIAP, February 2003.

[67] J. D. R. M. SILVIA CHIAPPA, *Eeg-based bci systems and idiap eeg database*, IDIAP-RR 64, IDIAP, 2003.

[68] T. A. STEPHENSON, *Conditional Gaussian mixtures*, IDIAP-RR 11, IDIAP, 2003.

[69] T. A. STEPHENSON, *Speech recognition with auxiliary information*, IDIAP-RR 28, IDIAP, 2003.

[70] V. TYAGI AND H. BOURLARD, *On multi-scale fourier transform analysis of speech signals*, IDIAP-RR 33, IDIAP, 2003.

[71] J.-P. T. V. POPOVICI, Y. RODRIGUEZ AND S. MARCEL, *On performance evaluation of face detection and localization algorithms*, IDIAP-RR 80, IDIAP, 2003.

[72] A. VINCIARELLI, *Noisy text categorization*, IDIAP-RR 61, IDIAP, 2003.

[73] A. VINCIARELLI, *Offline cursive handwriting: From word to text recognition*, IDIAP-RR 24, IDIAP, 2003.

[74] A. VINCIARELLI, *Effect of recognition errors on information retrieval performance*, IDIAP-RR 8, IDIAP, 2004.

[75] A. VINCIARELLI, *Noisy text categorization*, IDIAP-RR 3, IDIAP, 2004.

[76] A. VINCIARELLI AND S. BENGIO, *Transforming the feature vectors to improve HMM based cursive word recognition systems*, IDIAP-RR 32, IDIAP, 2002.

[77] A. VINCIARELLI, S. BENGIO, AND H. BUNKE, *Offline recognition of large vocabulary cursive handwritten text*, IDIAP-RR 01, IDIAP, 2003.

[78] K. WEBER, *Hmm mixtures (hmm2) for robust speech recognition*, IDIAP-RR 34, IDIAP, Martigny, Switzerland, 2003.

[79] S. B. Y. RODRIGUEZ, F. CARDINAUX AND J. MARIÉTHOZ, *Estimating the quality of face localization for face verification*, IDIAP-RR 07, IDIAP, 2004.

9.5 IDIAP Communications

[1] CONRAD SANDERSON, *Speech Processing & Text-Independent Automatic Person Verification*, IDIAP-Com 08, IDIAP, 2002.

[2] CONRAD SANDERSON, *The VidTIMIT Database*, IDIAP-Com 06, IDIAP, 2002.

[3] M. FLYNN AND P. WELLNER, *In search of a good bet*, IDIAP-COM 11, IDIAP, 2003.

[4] D. GRANGIER, A. VINCIARELLI, AND H. BOURLARD, *Information retrieval on noisy text*, IDIAP-COM 08, IDIAP, 2003.

[5] M. GUILLEMOT, J-M.ODOBEZ, AND D. GATICA-PEREZ, *Algorithms for video structuring*, IDIAP-COM 05, IDIAP, 2002.

[6] IDIAP, *Activity report 2001*, IDIAP-COM 01, IDIAP, 2002.

[7] IDIAP, *Activity report 2002*, IDIAP-COM 01, IDIAP, 2003.

[8] C. MARCEL, *Multimodal identity verification at idiap*, IDIAP-COM 04, IDIAP, 2003.

[9] I. MCCOWAN, D. GATICA-PEREZ, AND S. BENGIO, *Meeting data collection specifications*, IDIAP-COM 10, IDIAP, 2003.

[10] I. MCCOWAN AND D. MOORE, *Small microphone array: Algorithms and hardware*, IDIAP-COM 07, IDIAP, 2003.

[11] D. MOORE, *The idiap smart meeting room*, IDIAP-COM 07, IDIAP, 2002.

[12] D. MOORE, *Tode: A decoder for continuous speech recognition*, IDIAP-COM 09, IDIAP, 2002.

[13] A. C. MORRIS, *An information theoretic measure of sequence recognition performance*, IDIAP-COM 03, IDIAP, 2002.

[14] S. SARANGI, *Enhanced performance of multimodal biometric systems by confidence estimation*, IDIAP-COM 05, IDIAP, Martigny, Switzerland, 2003.

[15] J.-S. SENÉCAL, *Internship report : Summer 2003*, tech. rep.

[16] H. WANG AND S. BENGIO, *The mnist database of handwritten upper-case letters*, IDIAP-Com 04, IDIAP, 2002.

[17] H. WANG AND F. FORMAZ, *Idiap demonstration management*, IDIAP-Com 06, IDIAP, 2003.

[18] H. WANG, A. VINCIARELLI, AND F. FORMAZ, *Handwriting recognition demo*, IDIAP-Com 02, IDIAP, 2002.

9.6 IDIAP PhD Theses

[1] D. CHEN, *Text detection and recognition in images and video sequences*, PhD thesis, École Polytechnique Fédérale de Lausanne, Aug. 2003.

[2] T. STEPHENSON, *Speech Recognition with Auxiliary Variables*, PhD thesis, École Polytechnique Fédérale de Lausanne, 2003.

[3] A. VINCIARELLI, *Offline Cursive Handwriting Recognition: From Word to Text*, PhD thesis, University of Bern, 2003.

[4] K. WEBER, *HMM Mixture (HMM2) for Robust Speech Recognition*, PhD thesis, École Polytechnique Fédérale de Lausanne, 2003.