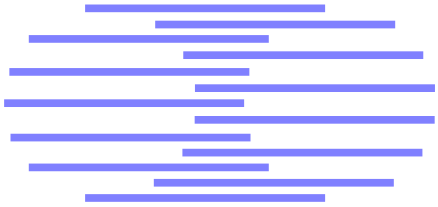
IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse



Activity Report 1998

(Avec Présentation Générale en Français)

(Mit Deutscher Allgemeiner Präsentation)

IDIAP–Com 99-01

March 99

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

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

IDIAP

Martigny - Valais - Suisse

Institut Dalle Molle d'Intelligence Artificielle Perceptive

MEMBERS

Supporting:

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

Affiliated:

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

FOUNDATION COUNCIL

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

BOARD OF DIRECTORS

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

SCIENTIFIC COMMITTEE

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

Inhaltsverzeichnis

1 *Présentation Générale de l'Institut* ii
1.1 Introduction . . . ii
1.2 Activités de recherche et développement . . . iii
1.3 Participation dans des projets de recherche nationaux et européens . . . vi
1.4 Collaborations avec d'autres organisations et sociétés . . . viii
1.5 Activités de formation et développement régional . . . ix
1.6 Publications . . . ix

1 *Allgemeine Präsentation des Instituts* xii
1.1 Einleitung . . . xii
1.2 Forschung und Entwicklung . . . xiii
1.3 Teilnahme an Nationalen und Europäischen Forschungsprojekten . . . xvi
1.4 Zusammenarbeit mit anderen Institutionen und Unternehmen . . . xviii
1.5 Ausbildung und Regionale Entwicklung . . . xviii
1.6 Veröffentlichungen . . . xix

1 General Overview of the Institute 1
1.1 Introduction . . . 1
1.2 Research and Development Activities . . . 2
1.3 Participation in National and European Community Research Projects . . . 5
1.4 Collaboration with other Organisations and Companies . . . 6
1.5 Training Activities and Regional Development . . . 7

2 Staff 10
2.1 Scientific Staff . . . 10
2.2 Visitors . . . 12
2.3 Students . . . 12
2.4 Administrative Staff . . . 12

3 Research Activities 14
3.1 Speech Processing Group . . . 14
3.1.1 Overview of the Speech Processing group activities . . . 14
3.1.2 Base Technology Tools . . . 16
3.1.3 Small / Medium Vocabulary Robust Speech Recognition . . . 19
3.1.4 Speaker Recognition . . . 23
3.1.5 Large Vocabulary Robust Speech Recognition . . . 26
3.1.6 Voice Thematic Indexing . . . 29
3.1.7 Prototyping and Spoken Language Resources . . . 30
3.1.8 Software Development . . . 32
3.1.9 Education . . . 32
3.2 Computer Vision Group . . . 34
3.2.1 Object Recognition . . . 34
3.2.2 Audio-Visual Person Verification . . . 35
3.2.3 Audio-Visual Speech Recognition . . . 40
3.2.4 Facial Expression Recognition . . . 41
3.2.5 X-Ray Image Sequence Analysis . . . 42
3.2.6 Document Analysis and Recognition . . . 43
3.3 Machine Learning Group . . . 46

3.3.1 Divide and learn . . . 46
3.3.2 Learn and understand what you learn . . . 49
3.3.3 Time series prediction and modeling . . . 51
3.3.4 Spatial data analysis . . . 52

4 Educational Activities **54**
4.1 Current Ph.D. Theses . . . 54
4.2 Ph.D. exams . . . 55
4.3 Student Projects . . . 56
4.4 Lectures . . . 57
4.5 Seminars . . . 58
4.6 Examinations . . . 59

5 Other Scientific Activities **62**
5.1 Editorship . . . 62
5.2 Scientific Committees Membership . . . 62
5.3 Organization of Conference . . . 63
5.4 Short term visits . . . 63

6 Events and Presentations **66**
6.1 Scientific Presentations . . . 66
6.2 Regional Presentations . . . 67

7 Publications (1997 and 1998) **68**
7.1 Books and Book Chapters . . . 68
7.2 Articles in International Journals . . . 68
7.3 Articles in Conference Proceedings . . . 69
7.4 IDIAP Research Reports . . . 72
7.5 IDIAP Communications . . . 73
7.6 Other Documents . . . 74

1 *Présentation Générale de l'Institut*

1.1 Introduction

L'Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP, `http://www.idiap.ch`) est un institut de recherche semi-privé à but non lucratif. Il a été créé en 1991 pour célébrer le 20ème anniversaire de la Fondation Dalle Molle, et représente le troisième centre de recherche initié par cette fondation, après l'ISSCO à Genève (`http://www.issco.ch`) et l'IDSIA à Lugano (`http://www.idsia.ch`).

En Novembre 1996, et comme convenu lors de sa création, l'IDIAP a acquis le statut de fondation de recherche (Fondation IDIAP), désormais indépendante de la Fondation Dalle Molle, et dont les fondateurs sont la ville de Martigny, l'État du Valais, l'École Polytechnique Fédérale de Lausanne (EPFL), l'Université de Genève et Swisscom.

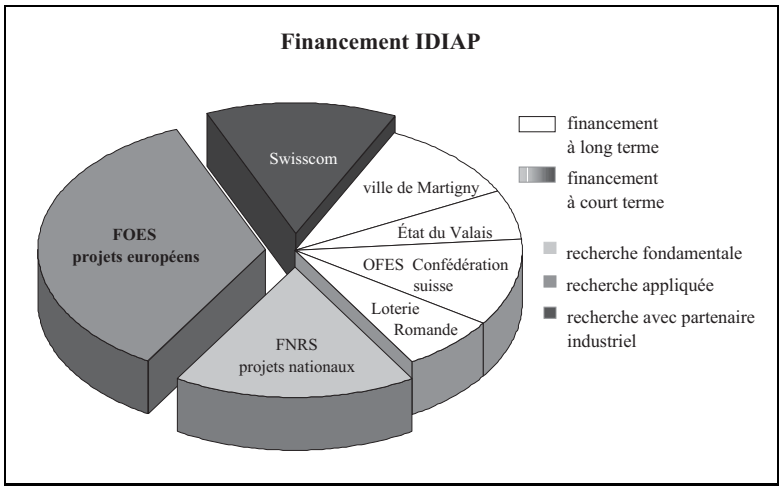


Fig. 1 – *Distribution relative du financement de l'IDIAP en 1998.*

Aujourd'hui, l'IDIAP est principalement financé par un support à long terme de la confédération suisse (Office Fédéral de l'Éducation et de la Science, OFES), de l'État du Valais, de la ville de Martigny et de Swisscom. La Loterie Romande supporte également nos efforts de recherche au travers d'un subside annuel. En plus de ces financements de base, l'IDIAP bénéficie de plusieurs projets financés par le fonds national suisse de la recherche scientifique (FNSRS) pour de la recherche de base (surtout des étudiants doctorants), ainsi que de l'OFES dans le cadre de projets européens. La distribution relative des différentes sources de financement de l'IDIAP est représentée à la Figure 1.

Depuis quelques années, l'IDIAP emploie entre 25 et 30 scientifiques, composés essentiellement de personnel permanent, de chercheurs post-doctorat, d'ingénieurs doctorants, et de visiteurs à court ou moyen terme. Au début 1999, l'IDIAP emploie 30 personnes, dont 10 scientifiques seniors, 5 ingénieurs, 13 étudiants doctorants et 2 secrétaires.

La structure de gestion de l'IDIAP est représentée à la Figure 2 et est composée d'un conseil de fondation, d'un comité de direction, et d'un conseil scientifique (conseillant le comité de direction). Il est également prévu de mettre sur pied un petit comité de relations économiques, lequel sera responsable de la communciation des résultats au monde industriel, tout en signalant à l'IDIAP les nouvelles opportunités de recherche ayant un intérêt particulier pour l'industrie.

Les activités de l'IDIAP peuvent se répartir selon différentes catégories: les activités de recherche et développement, la participation à de nombreux projets de recherche européens et nationaux, les

collaborations avec diverses organisations et sociétés, et les activités d'enseignement et de formation. La mission de l'IDIAP consiste donc en:

- La poursuite d'activités de recherche fondamentale et appliquée, dans le but de transfert technologique à moyen et long terme.
- L'enseignement et la formation.

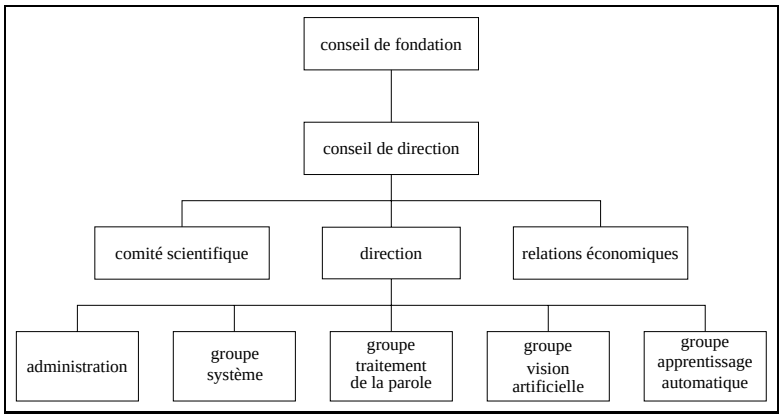


Fig. 2 – *Structure générale de l'IDIAP.*

1998 a été une bonne année pour notre institut, et les activités de l'IDIAP ont été des plus florissantes. Le nombre de projets nationaux et internationaux, ainsi que le partenariat avec les institutions académiques, n'ont cessé de s'accroître. De plus, grâce au soutien continu de nos institutions, et aux compétences élevées de notre personnel, motivé par un travail d'excellente qualité, l'IDIAP est maintenant reconnu comme un partenaire de haut niveau dans ses domaines de compétence (traitement automatique de la parole, vision par ordinateur, apprentissage automatique). Il nous reste maintenant à poursuivre nos activités de recherche et développement dans ces domaines d'expertise, tout en favorisant le transfert technologique au travers de partenariats industriels (y compris avec VOXCom, une société "spin-off" de l'IDIAP démarrée en juillet 1998 – voir fin de la section 1.3).

1.2 Activités de recherche et développement

Les activités de recherche de l'IDIAP sont orientées vers les **interactions multimodales** et se répartissent actuellement en trois groupes: traitement de la parole, vision artificielle et apprentissage automatique. Dans ces trois domaines complémentaires, et se focalisant sur quelques axes bien définis, l'IDIAP poursuit des travaux de recherche à moyen et long terme et développe des systèmes prototypes (dans le but de valider les résultats de recherche). Pour 1998, les activités de recherche et de développement dans chacun de ces trois groupes peuvent se résumer comme suit.

- **Traitement automatique de la parole, incluant tous les aspects de la reconnaissance automatique de la parole et de la vérification du locuteur**.
  Cette activité comprend le développement et l'évaluation de systèmes avancés de reconnaissance automatique de la parole par ordinateur, ainsi que des systèmes représentant l'état de l'art (et couvrant les systèmes à petits et grands vocabulaires, dépendants ou indépendants du locuteur, reconnaissant des mots isolés ou de la parole continue, ou capables de détecter des mots clés). Bien que l'IDIAP se focalise surtout sur les environnements téléphoniques (étant donné notre collaboration avec Swisscom), nos travaux sont également testés sur des applications à entrée microphone. Les activités de recherche actuelles se focalisent surtout sur l'amélioration des modèles d'unités de parole et sur le traitement du signal de façon à améliorer la robustesse des systèmes au bruit et aux styles d'élocution. Ceci met notamment en jeu le développement de techniques

avancées d'adaptation automatique, l'amélioration des modèles de Markov cachés (HMM) et des systèmes hybrides utilisant les modèles HMM conjointement avec les réseaux de neurones artificiels, ainsi que des techniques avancées de traitement en sous-bande et multi-canaux (comme initié par l'IDIAP, en collaboration avec la Faculté Polytechnique de Mons en Belgique, le " International Computer Science Institute" de Berkeley, USA, et le "Oregon Graduate Institute" de Portland, USA). Des systèmes de reconnaissance de parole continue et grands lexiques, mettant en oeuvre des dictionnaires et règles de prononciations complexes, ainsi que des contraintes grammaticales avancées, sont également développés et testés.

Comme décrit par la suite, le groupe de traitement de la parole de l'IDIAP est engagé dans de nombreux projets nationaux et internationaux (tels que les projets européens ESPRIT, ACTS, COST, et TMR).

En vérification automatique de locuteur, la plupart des activités de recherche se focalisent sur l'amélioration de l'état de l'art, ainsi que sur le développement de solutions innovatrices combinant des stratégies concurrentes ou complémentaires. Ces deux dernières années, l'IDIAP a participé aux évaluations internationales NIST (National Institute of Standards and Technology, USA), démontrant ainsi que sa technologie était particulièrement compétitive.

Les principales applications et systèmes prototypes qui ont été développés et testés jusqu'à présent à l'IDIAP comprennent: les serveurs vocaux interactifs (permettant, par exemple, l'accès vocal à des bases de données), les répertoires vocaux personnalisés et majordomes vocaux (par exemple, PABX avec reconnaissance de la parole), les applications cartes téléphoniques (mettant en oeuvre la composition vocale de numéros de cartes ou numéros téléphoniques, ainsi que la vérification du locuteur), l'indexation automatique et recherche de documents audio, et la vérification multimodale (parole et vision) de l'identité de personnes.

Finalement, afin de faciliter la recherche et le développement de systèmes multi-langues, l'IDIAP est activement engagé dans l'enregistrement, l'étiquetage, et la maintenance de grandes bases de données de parole. Cette activité s'inscrit soit dans le cadre de notre collaboration avec Swisscom (Polyphone et données GSM), soit comme partenaire d'un large effort européen pour le développement de grandes bases de données multi-langues.

– **Vision par ordinateur, incluant la reconnaissance d'objets, l'analyse du mouvement, la fusion de modes, et la reconnaissance de documents**.

La vision par ordinateur en général traite de l'interprétation et de l'analyse automatique de scènes visuelles. Bien que ce domaine d'activités soit très vaste, la stratégie du groupe est de se focaliser sur des domaines de recherche ciblés sur des applications bien précises, telles que le domaine des interfaces multimodaux, les accès sécurisés, ainsi que la gestion et l'accès aux informations multimédia.

À travers ses activités dans différents projets, le groupe Vision a maintenant acquis une considérable expérience dans les domaines de la détection et la reconnaissance d'objets, l'analyse et la représentation de formes, l'analyse et la reconnaissance de mouvements, la fusion de modalités, ainsi que l'analyse et la reconnaissance de documents.

Beaucoup de ces développements bénéficient de la collaboration avec le groupe de reconnaissance de la parole (par exemple, pour la reconnaissance de mouvement par HMM, ou la reconnaissance multimodale utilisant le traitement multi-canaux récemment proposé par le groupe parole), ainsi que le groupe d'apprentissage automatique (par exemple, dans le cas de la combinaison de classificateurs).

Les résultats de recherche obtenus ici ont notamment conduit à de nouvelles technologies et applications. La technique de détection d'objets a été appliquée au problème de détection de visages et a été intégrée dans un système prototype fonctionnant en temps réel. Un algorithme de suivi du mouvement des lèvres a été développé et, en combinaison avec les techniques de reconnaissance de mouvement, a conduit au développement de systèmes de reconnaissance visuelle de la parole, ainsi qu'à la vérification visuelle de l'identité de personnes. Cette approche a également été combinée avec l'analyse acoustique du signal de parole pour conduire à des

systèmes de reconnaissance audio-visuelle de la parole et de la vérification audio-visuelle d'identité. Dans ce cadre, plusieurs méthodes de fusion de données ont été étudiées et testées sur des systèmes multimodaux de vérification d'identité. Un de ces systèmes a notamment été intégré par Cerberus AG et Ibermatica S.A. dans une application prototype. Nos travaux en analyse d'images aux rayons-X ont conduit à une méthode d'extraction de paramètres articulatoires dans des séquences d'images aux rayons-X. Finalement, plusieurs approches pour l'analyse de documents imprimés ou manuscrits ont été étudiées dans le groupe.

– **Apprentissage automatique, incluant la reconnaissance de formes, l'analyse de données et l'extraction de connaissances**.

  Le principal but de ce groupe est de maintenir une expertise forte dans différentes disciplines avancées ayant été identifiées comme étant d'intérêt direct pour les travaux présents et futurs de l'IDIAP. Ces disciplines couvrent des domaines très variés tels que l'apprentissage bayésiens, les réseaux de neurones artificiels, les arbres de décisions, les machines à vecteurs supports, et l'analyse logique de données.

  La collaboration entre ces différentes techniques d'apprentissage et les applications spécifiques (telles que la classification de formes et la reconnaissance de la parole) étudiées à l'IDIAP est généralement fructueuse et a déjà conduit au développement d'approches originales et résultats intéressants. Il est cependant clair que ce type de recherche demande un effort important pour adapter des méthodes générales à des problèmes spécifiques, tels que le traitement de grandes bases de données bruitées (bases de données parole). Plus particulièrement, un effort important est investi dans la décomposition de larges problèmes complexes en un ensemble de sous-problèmes plus simples.

  Finalement, en vue d'identifier de nouveaux domaines de recherche prometteurs, l'expertise du groupe est également exploitée dans des activités plus prospectives telles que la prédiction de séries temporelles (avec applications, par exemple, en prédiction de marchés boursiers ou de risques d'avalanches) et le développement de systèmes d'aide aux diagnostics.

– **Groupe système, incluant aussi la gestion des bases de données et le développement de systèmes prototypes**.

  Les trois groupes de recherche décrits ci-dessus sont secondés par un petit Groupe Système également responsable de la gestion des nombreuses bases de données et du développement de systèmes prototypes, et travaillant en étroite collaboration avec les groupes de recherche dans le cas de projets plus orientés vers les applications.

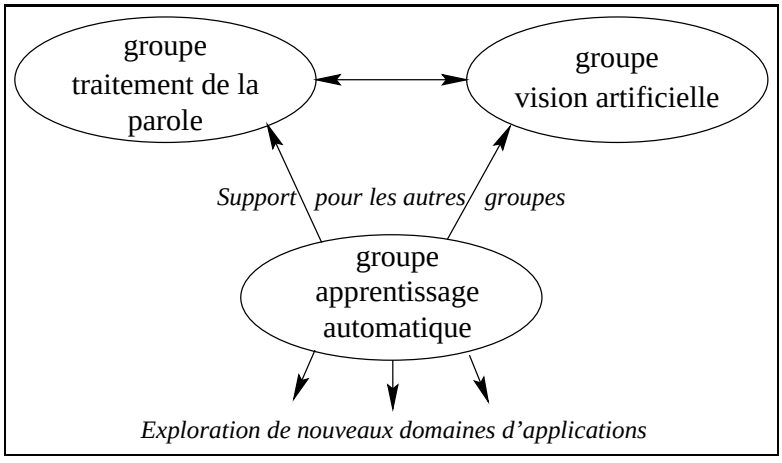


Fig. 3 – *Interactions entre les trois groupes de recherche.*

Comme brièvement décrit ci-dessus, et comme illustré à la Figure 3, les activités des trois groupes de recherche ont été définies de façon à être complémentaires et à favoriser une collaboration active entre les différents thèmes de recherche.

Alors que le traitement de la parole et de la vision par ordinateur sont souvent complémentaires dans des applications multimodales, ces disciplines sont aussi souvent basées sur des théories et outils mathématiques communs, et bénéficient dès lors de cette interaction. Par exemple, des développements récents en reconnaissance d'écriture manuscrite ont utilisé les modèles de Markov cachés initialement développés en reconnaissance de la parole. De même, des progrès récents en traitement multi-canaux (initié par l'IDIAP, en collaboration avec d'autres laboratoires tels que l'ICSI de Berkeley, USA, et la FPMs de Mons, Belgique) seront exploités dans les groupes parole et vision, et profiteront directement au développement de systèmes multimodaux (comme récemment démontré par l'IDIAP par des travaux préliminaires en reconnaissance de la parole audio-visuelle).

Le groupe d'apprentissage automatique apporte le support théorique supplémentaire aux deux autres groupes (plus focalisés sur les applications) en étudiant de nouvelles technologies qui sont communes et utiles à la fois au traitement de la parole et de la vision, ainsi qu'au traitement multimodal. Par exemple, en 1998, les nouvelles méthodes proposées pour la décomposition de larges problèmes d'apprentissage en sous-problèmes ont été appliquées avec succès au problème de vérification automatique du locuteur. De plus, la recherche sur les différentes façons de recombiner les sous-modules résultant de cette décomposition a contribué au problème de la fusion des experts multimodaux. Finalement, la deuxième mission du groupe d'apprentissage automatique est d'identifier et d'étudier de nouveaux domaines d'applications qui pourraient bénéficier des technologies disponibles à l'IDIAP (et d'abord développées dans le cadre de la parole et de la vision) et qui pourraient devenir de futures activités importantes pour l'IDIAP (comme, par exemple, la prédiction de séries temporelles). Dans ce cas, la recherche sera souvent effectuée en collaboration avec d'autres institutions ayant une plus grande expertise dans le nouveau domaine.

1.3 Participation dans des projets de recherche nationaux et européens

Les activités de l'IDIAP dans le cadre de projets de recherche nationaux et internationaux (surtout des projets émanant de la Commission des Communautés Européennes) ont été particulièrement intenses durant ces quelques dernières années, et l'IDIAP a souvent joué un rôle clé dans la conception et la réalisation de ces projets. Au terme de l'année 1998, on dénombre 8 projets du Fonds National Suisse pour la Recherche Scientifique et l'IDIAP est partenaire dans 8 projets européens. De plus, à travers sa collaboration avec Swisscom, l'IDIAP participe à un important projet industriel.

En 1998, pour ce qui concerne les projets financés par les Communautés Européennes (et par l'OFES pour les partenaires suisses) dans le cadre du 4ème Programme Cadre pour la Recherche et la Technologie, l'IDIAP était un partenaire actif de nombreux projets, incluant:

- Deux projets dans le domaine "télématique":
  - *SpeechDat*, visant à produire, standardiser et évaluer de très grandes bases de données parole couvrant la plupart des langues européennes. Ce projet a notamment permis à l'IDIAP de développer des bases de données pour le français et l'allemand parlés en Suisse.
  - *PIoneering Caller Authentication for Secure Service Operation (PICASSO)*, concernant l'utilisation de systèmes de vérification automatique du locuteur dans des applications de cartes téléphoniques ou de majordomes vocaux personnalisés.
- Deux projets ESPRIT pour la recherche à long terme:
  - *THematic Indexing of Spoken Language (THISL)*, traitant de l'indexation automatique et de l'accès vocal de documents sonores, et plus particulièrement des nouvelles (radio/TV) de la BBC.
  - *REcognition of Speech by Partial Information TEchniques (RESPITE)*, concernant l'étude de nouvelles techniques de reconnaissance automatique de la parole qui seraient plus robustes aux bruits et corruptions divers, et de leur déploiement dans deux domaines d'ap-

plication (téléphones portables et systèmes embarqués). Bien que ce projet ait été accepté en 1998, il a officiellement démarré le 1er janvier 1999.

- Dans le cadre du programme européen *Training and Mobility of Researchers (TMR)*, ayant pour but de promouvoir la mobilité internationale des chercheurs, l'IDIAP est un des partenaires clés dans le projet SPHEAR (SPeech, HEAring and Recognition) dont l'objectif est d'acquérir une meilleure compréhension du système auditif humain, et d'intégrer certaines de ses propriétés dans les systèmes de reconnaissance afin d'en améliorer la robustesse.
- Deux projets COST (European Cooperation in the field of Scientific and Technical Research): COST249 concernant la reconnaissance automatique de la parole sur ligne téléphonique, et COST250 relatif à la vérification automatique du locuteur sur ligne téléphonique. En comparaison des autres partenaires européens, il est important de noter ici que l'IDIAP a bénéficié d'un support important de l'OFES dans le cadre de ces deux projets COST, ce qui lui a permis d'augmenter de façon significative son expertise dans les domaines de la reconnaissance de la parole et du locuteur, et d'initier de nouveaux contacts européens et industriels.
- Un projet ACTS (Advanced Communications Technologies and Services) sur la vérification multimodale dans les applications d'accès sécurisé aux télé-services et bâtiments.
- Finalement, dans le cadre du programme Socrates/Erasmus, l'IDIAP est le seul partenaire suisse dans un projet européen dont le but est de mettre en place un programme de master européen en technologie du langage et de la parole. Le contenu des cours communs aux différents partenaires a été défini, donnant la possibilité aux étudiants de poursuivre les cours et les projets dans différentes institutions. En 1998, l'IDIAP a invité l'EPFL à se joindre à cette initiative.

Au niveau national, l'IDIAP est engagé dans plusieurs projets du fonds national suisse de la recherche scientifique (FNSRS) finançant principalement des étudiants doctorants (enregistrés à l'EPFL, l'Université de Genève, ou l'Université de Lausanne).

- AV-COM: *Audio-Visual Combination*, concernant la combinaison audio-visuelle et la reconnaissance multimodale (octroyé en 1997, pour de l'équipement spécialisé).
- MULTICHAN: *Non-stationary multichannel signal processing*, relatif au traitement de signaux multi-canaux non stationnaires appliqué à la reconnaissance de la parole (plus particulièrement les approches multi-bandes et multi-résolution).
- SV-UCP: *Speaker Verification based on User-Customized Password*, pour la vérification automatique du locuteur sur base de mots de passe facilement défini par l'utilisateur.
- INSPECT: *INtegrating SPeech ConsTraints for enhanced recognition systems*, concernant l'intégration efficace de contraintes syntaxiques et sémantiques complexes dans les systèmes de reconnaissance de la parole, en collaboration avec l'EPFL (Dr Martin Rajman, DI/LIA).
- BN-ASR: *Modelling the hidden dynamic structure of speech production in a unified framework for robust automatic speech recognition*, étudiant les possibilités d'utilisation des réseaux bayésiens comme nouvelle méthode de reconnaissance de la parole, afin de permettre une meilleure modélisation des structures dynamiques du signal.
- ARTIST: *Articulatory Representation Towards Improved Speech Technology*, étudiant les possibilités d'extraire et d'utiliser les paramètres articulatoires dans les systèmes de traitement de la parole.
- FaceX: *Robust Facial Expression Recognition through Temporal and Appearance-based Models*, ayant pour objectif de développer des méthodes robustes permettant l'analyse visuelle et la reconnaissance d'expressions du visage dans les séquences d'images.
- GLAD: *Generalization of "Logical Analysis of Data" techniques*, étudiant et généralisant une nouvelle méthode de classification de données binaires, et permettant également d'interpréter les règles extraites automatiquement.
- *Compact hardware-friendly neural networks*, développant des méthodes de classification basées sur les réseaux de neurones artificiels, et plus particulièrement sur les mélanges d'experts.

- SEPHYR: *time SEries Prediction with Hybrid maRkov models*, étudiant et comparant différentes méthodes (réseaux de neurones, modèles de Markov cachés) pour la prédiction de séries temporelles (prédiction du marché boursier, prédiction des risques d'avalanches).
- CARTANN *Cartography by Artificial Neural Networks*, en collaboration avec l'Université de Lausanne (Prof. Michel Maignan, département des sciences de la terre, institut de géostatistique), ce projet a pour but d'explorer les potentialités des réseaux de neurones (et autres approches statistiques) à traiter certains problèmes de géostatistiques (avec applications, par exemple, dans l'étude de l'évolution de la pollution autour du Lac Léman).

Finalement, bien que l'IDIAP ait été engagé par le passé dans un projet CTI (Commission pour la Technologie et l'Innovation), 1998 a été une année de transition pendant laquelle VOXCom, une société "spin-off" de l'IDIAP, a été initiée. Autour de VOXCom, un nouveau projet CTI (InfoVOX, mettant en oeuvre l'IDIAP, l'EPFL, Swisscom, VOXCom S.A. et Omedia S.A.) a été défini, soumis, et accepté par la CTI avec une date de démarrage fixée au 1er mars 1999. Un des buts du projet InfoVOX est de poursuivre les recherches et les efforts de développement dans le domaine des serveurs vocaux interactifs, avec développement d'applications de téléphonie informatique. L'application générique visée dans ce projet concerne les systèmes vocaux interactifs permettant l'accès par la parole aux systèmes d'information. Plus particulièrement, InfoVOX se focalisera sur l'accès aux bases de données internet, et un système utilisant une interface vocale à la page web de l'office de tourisme de Martigny (`http://www.martigny.ch`) sera testé.

La plupart des projets mentionnés ci-dessus sont discutés plus en détail dans le présent rapport d'activités.

1.4 Collaborations avec d'autres organisations et sociétés

Depuis ces dernières années, l'IDIAP a maintenu des contacts étroits avec différents instituts de recherche, universités et industries partageant les mêmes domaines d'intérêt. En général, ces contacts résultent du suivi de certains projets communs, ou sont basés sur des relations personnelles à long terme. Parmi ces contacts, nous pouvons mentionner ici:

- La collaboration active entre l'IDIAP et Swisscom (dont une brève description est donnée dans le présent rapport d'activités).
- Le partenariat avec les institutions académiques telles que l'EPFL (où Hervé Bourlard est également Professeur), l'Université de Genève, et l'IMT (Université de Neuchatel). Plusieurs projets de recherche communs entre l'IDIAP et l'EPFL sont actuellement en cours, et plusieurs étudiants doctorants de l'IDIAP sont inscrits à l'EPFL.
- Initiés grâce aux projets européens, de bons contacts ont été établis entre l'IDIAP et de nombreuses sociétés, dont: Cerberus (CH), BBC (UK), Daimler-Benz (D), Thomson (F), Matra Nortel (F), Ibermatica (E), ainsi qu'avec de nombreuses universités étrangères, dont: l'Université de Cambridge (UK), l'Université de Sheffield (UK), la Faculté Polytechnique de Mons (BE), et l'Université de Surrey (UK).
- Par des contacts personnels et des échanges réguliers d'informations, l'IDIAP collabore également activement avec l'Université de Rutgers (RUTCOR, USA) et l'"International Computer Science Institute" (ICSI, Berkeley, USA), incluant notamment l'échange d'étudiants.
- Plus récemment, et dans le cadre d'un projet financé par la Fondation Catalyst (USA), nous avons démarré un projet de 4 ans en collaboration avec l'université Johns Hopkins University (Baltimore, USA) et le "Indian Institute of Technology" (Delhi, Inde) sur l'étude et le développement d'un circuit analogique à très basse puissance capable de faire la reconnaissance de parole continue.

1.5 Activités de formation et développement régional

En plus des activités de recherche et développement, l'IDIAP a également deux autres missions importantes: la formation et le transfert technologique vers l'industrie.

Formation et supervision d'étudiants doctorants (souvent enregistrés à l'EPFL, l'Université de Genève ou l'Université de Lausanne), de chercheurs post-doctorat, et de visiteurs à moyen et long terme originaires des milieux académiques (incluant les ETS) et industriels. Un exemple typique de cette activité concerne l'engagement de l'IDIAP dans un projet européen (Socrates/Erasmus) pour la mise en place de Masters Européens en technologie du langage et de la parole. Les premiers tests (cours) de ce programme de Masters devraient démarrer en 1999.

L'IDIAP est donc très actif dans la formation de chercheurs et d'ingénieurs. En plus de ses 13 thèses de doctorat actuellement en cours, une thèse a été achevée durant l'année 1998, sous le label de l'École Polytechnique Fédérale de Lausanne (EPFL). Durant l'année écoulée, 7 étudiants, pour la plupart futurs ingénieurs, ont effectués leur travail de fin d'étude à l'IDIAP. Tout au long de l'année, le Prof. Hervé Bourlard a enseigné à l'école pré-doctorale du département d'informatique ainsi que de la section de systèmes de communications de l'EPFL. Divers membres de l'IDIAP ont donnés des séminaires tout au long de l'année à l'EPFL et à l'ETHZ, dans diverses universités: Genève, Gènes (IT), Surrey (UK), Londres (UK). Le Prof. Hervé Bourlard fut membre du comité de plusieurs thèses de doctorat. Finalement, l'IDIAP accueille régulièrement des étudiants ingénieurs préparant leur thèse de fin d'études (thèse de diplôme) et venant de l'EPFL, Eurecom (F), ENST (F), ainsi que de l'École d'Ingénieurs du Valais (EIV).

Transfert technologique et support industriel, avec deux motivations: (1) permettre aux industries de se tenir à jour avec la technologie (étant donné qu'il est maintenant souvent trop onéreux, même pour les plus grosses sociétés, de maintenir leurs propres compétences dans tous les domaines importants), et (2) permettre à l'IDIAP de développer et tester des systèmes prototypes orientés vers des applications d'intérêt direct pour nos sponsors (par exemple, Swisscom). Dans ce cadre, l'IDIAP doit également être capable de fournir des analyses qualitatives et quantitatives, d'améliorer l'applicabilité des technologies de bases, d'intégrer ces technologies dans des systèmes pilotes, et de s'engager dans du transfert actif de technologie.

Au travers d'une collaboration industrielle ouverte et intensive, l'IDIAP entend jouer un rôle important dans le développement économique de l'État du Valais. Dans ce cadre, VOXCom S.A. a démarré en juillet 1998 sous l'impulsion de l'IDIAP et de la ville de Martigny. En partant des solutions disponibles à l'IDIAP, la priorité première de VOXCom est de développer et d'intégrer des produits logiciels et des services répondant aux besoins spécialisés des sociétés se tournant vers la téléphonie informatique et les serveurs vocaux interactifs.

Finalement, l'IDIAP est régulièrement engagé dans l'organisation d'événements scientifiques tels que:

- La première conférence internationale sur "Audio-and Video-based Biometric Person Authentication" (AVBPA), organisée par l'IDIAP, s'est tenue à Crans-Montana du 12 au 14 mars 1997.
- Du 15 au 19 juin 1998 se sont tenues au CERM à Martigny les *XXIIemes Journées d'Étude sur la Parole* (JEP'98), organisée par l'IDIAP et le GFCP (Groupe Francophone de la Communication Parlée).

1.6 Publications

La valeur d'un institut de recherche scientifique est essentiellement jaugée à ses publications (nombre, mais surtout qualité). Pour 1997 et 1998, les publications de l'IDIAP (énumérées plus en détails à la fin du présent rapport d'activités) sont les suivantes:

- 2 livres (en préparation);
- 7 chapitres de livre;
- 10 articles dans des revues internationales;

- 45 articles dans des conférences internationales;
- 28 rapports scientifiques internes.

1 *Allgemeine Präsentation des Instituts*

1.1 Einleitung

Das Dalle Molle Institut für Perzeptive Künstliche Intelligenz (IDIAP, "Institut Dalle Molle d'Intelligence Artificielle Perceptive", `http://www.idiap.ch`) ist ein halbprivates gemeinnütziges Forschungsinstitut, das im Jahre 1991 anlässlich des 20-jährigen Jubiläums der Dalle Molle Stiftung gegründet wurde. Nach ISSCO in Genf (`http://www.issco.ch`) und IDSIA in Lugano (`http://www.idsia.ch`) ist es das dritte Forschungsinstitut, das von der Dalle Molle Stiftung initiiert wurde.

Wie anlässlich der Gründung des Instituts ursprünglich geplant, erlangte IDIAP im November 1996 den Status einer Forschungsstiftung "Stiftung IDIAP". IDIAP wurde damit unabhängig von der Dalle Molle Stiftung. Die Gründer der "Stiftung IDIAP" sind die Stadt Martigny, der Kanton Wallis, die Eidgenössische Technische Hochschule Lausanne (EPFL), die Universität Genf und Swisscom.

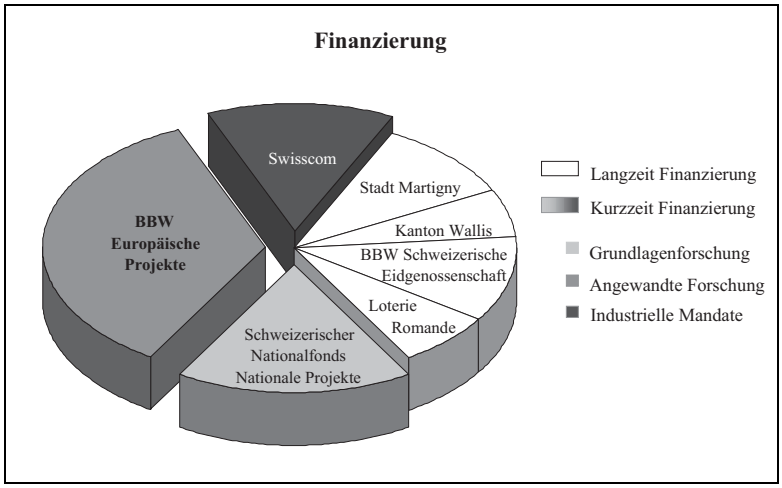


Abbildung 4: *Relative Verteilung der Finanzierung von IDIAP im Jahre 1998.*

Langfristige finanzielle Unterstützung erhält IDIAP heutzutage hauptsächlich von der Schweizerischen Eidgenossenschaft (Bundesamt für Bildung und Wissenschaft, BBW), dem Kanton Wallis, der Stadt Martigny und Swisscom. Ausserdem fördert die "Loterie Romande" unsere Forschungsbemühungen mit jährlichen Beiträgen. Zusätzlich erhält IDIAP wesentliche Forschungbeiträge vom Schweizerischen Nationalfonds (SNF) für Projekte der Grundlagenforschung und vom BBW für europäische Projekte. Die relative Verteilung der Finanzierung von IDIAP zeigt Abbildung 4.

In den letzten Jahren waren im Durchschnitt 25-30 Wissenschaftler am IDIAP tätig, die sich zusammensetzen aus fest angestellten Wissenschaftlern, Forschungsassistenten, Doktoranden und Gastwissenschaftlern.

Die Managementstruktur von IDIAP ist in Abbildung 5 dargestellt. Sie setzt sich zusammen aus dem Stiftungsrat, dem Direktionskomitee und dem wissenschaftlichen Komitee (berät die Direktion). Es ist auch beabsichtigt, ein Komitee für wirtschaftliche Relationen ins Leben zu rufen, das zur Aufgabe hat, IDIAPs Forschungsresultate in der Industrie bekanntzumachen und das verantwortlich sein wird, IDIAP neue Forschungsmöglichkeiten, die von speziellem Interesse für die Industrie sind, aufzuzeigen.

Unsere Tätigkeitsfelder beinhalten Forschung und Entwicklung, Teilnahme an europäischen und nationalen Projekten, Zusammenarbeit mit Organisationen und Unternehmen, als auch Lehre und Ausbildung. IDIAPs Aufgaben bestehen deshalb aus:

- Grundlagen- und angewandter Forschung, die sich in einem mittleren bis längeren Zeitrahmen wirtschaftlich bezahlt machen.
- Lehre und Ausbildung.

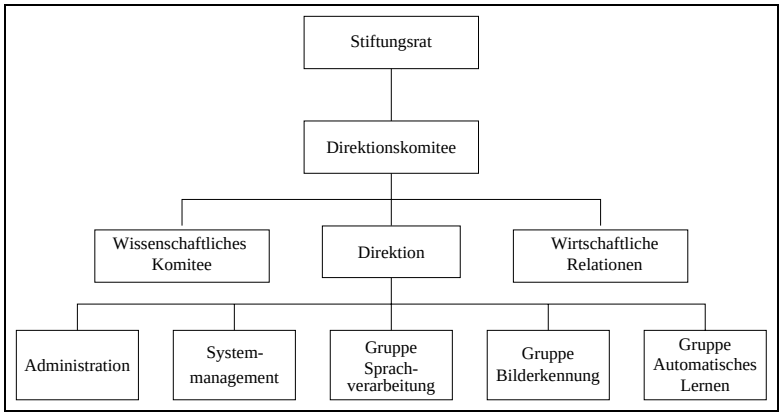


Abbildung 5: *Die Unternehmensstruktur von IDIAP.*

Das Jahr 1998 war ein erfolgreiches Jahr für unser Institut so dass eine positive Bilanz gezogen werden kann: die Anzahl nationaler und internationaler Projekte als auch die Partnerschaften mit akademischen Institutionen sind gestiegen. Ferner ist IDIAP dank der kontinuierlichen Unterstützung unserer Behörden und unseres kompetenten und motivierten Personals, ein begehrter Partner auf den Gebieten Sprachverarbeitung, Bilderkennung und Automatisches Lernen. Es ist nun unsere Aufgabe, uns weiter auf die Forschung und Entwicklung auf den genannten Gebieten zu konzentrieren und zugleich Forschungsresultate in industrielle Partnerschaften einfliessen zu lassen. Dies betrifft auch VOXCom, eine Spin-Off-Firma von IDIAP, gegründet im Juli 1998 – siehe weiter unten.

1.2 Forschung und Entwicklung

Unsere Hauptforschungsaktivitäten im Jahre 1998 konzentrierten sich auf mittel- bis langfristige Ziele und werden im Detail im vorliegenden Jahresbericht beschrieben. Wir betreiben Grundlagenforschung und entwickeln Prototypen (um Forschungsresultate zu validieren) im Bereich **multimodal interaction** der sich auf drei Forschungsgruppen verteilt :

- **Sprachverarbeitung, einschliesslich automatische Spracherkennung und Sprecherverifizierung**.

  Dies umfasst die Entwicklung und das Testen von fortgeschrittenen und dem Stand der Technik entsprechenden Spracherkennungssystemen (mit kleinem bis grossem Wortschatz, entweder sprecherabhängig oder sprecherunabhängig, von einzelnen Wörtern bis zu kontinuierlicher Sprache, als auch das Auffinden von Schlüsselwörtern / Schlüsselausdrücken). Während das Schwergewicht auf Telefonsprache (im Rahmen unserer Zusammenarbeit mit Swisscom) liegt, können unsere Systeme auch auf Mikrofoneingänge angewandt werden. Laufende Forschungsaktivitäten konzentrieren sich hauptsächlich auf die Verbesserung von verschiedenen Sprachmodelleinheiten bezüglich der Robustheit in Gegenwart von Hintergrundgeräuschen und Sprachvariationen.

Dies beinhaltet unter anderem Techniken für die Echtzeitadaptierung, Weiterentwicklungen der "Hidden Markovmodelle" (HMM) und der Hybridsysteme, die HMM zusammen mit neuronalen Netzwerken anwenden, als auch fortgeschrittene Forschung in Sub-Band und Multi-Stream Verarbeitung (eine Technik bei deren Entwicklung IDIAP zusammen mit der Faculté Polytechnique de Mons in Belgien, dem International Computer Science Institute in Berkeley, USA, und dem Oregon Graduate Institute in Portland, USA, massgeblich beteiligt war). Ebenso werden Spracherkennungssysteme mit grossem Wortschatz, die komplexe Aussprachewörterbücher und Grammatik benutzen, entwickelt und getestet. IDIAPs Sprachverarbeitungsgruppe beteiligt sich, wie weiter unten beschrieben, bei zahlreichen nationalen und internationalen Projekten (ESPRIT, ACTS, COST, TMR).

Bisher hat sich unsere Forschung in Sprecherverifizierung auf die Verbesserung von Algorithmen konzentriert, die dem gegenwärtigen Stand der Technik entsprechen sowie der Entwicklung von innovativen Lösungen, die konkurrierende und/oder sich ergänzende Strategien kombinieren. IDIAP hat kürzlich an der von NIST (National Institute of Standards and Technology, USA) organisierten internationalen Bewertung teilgenommen und gezeigt, dass unsere Technologie auf diesem Gebiet eine Spitzenposition einnimmt.

Die wichtigsten Anwendungen und Prototypen, die bisher entwickelt und getestet wurden umfassen: Sprach-Server (z.B. für Zugriff auf zentrale Datenbanken), Persönliche Anrufassistenten, Telefonkartenanwendungen (einschliesslich sprachgesteuerter Nummernwahl und Sprecherverifizierung), automatische Audioindizierung und Abfrage sowie multimodale Systeme zur Benützerverifizierung. Schliesslich ist IDIAP aktiv beteiligt am Sammeln und Management von Sprachdatenbanken zur Unterstützung der Forschung und Entwicklung von Mehrfachsprachsystemen, was hauptsächlich im Rahmen unseres Vertrages mit Swisscom (Polyphon und GSM Daten) als auch innerhalb eines EU Projektes erfolgt.

- **Bilderkennung, einschliesslich Objekterkennung, Bewegungsanalyse, Sensor Fusion und Dokumenterkennung.**

  Bilderkennung handelt im Allgemeinen von der Analyse und der Interpretation von visuellen Szenen. Die Strategie unserer Gruppe ist es, Forschungsthemen anzugehen, die von potentiellen Applikationen bestimmt werden. Sie hat zum Ziel, die Entwicklung neuer Technologien auf den Gebieten der multimodalen Schnittstellen, Zutrittskontrollen und Informationsmanagement zu fördern.

  Durch unsere Aktivitäten in verschiedenen Projekten hat die Gruppe Sachverstand erlangt in den Gebieten Objektsuche und -erkennung, Formanalyse, Bewegungsanalyse und -erkennung, Sensor Fusion, als auch Dokumentanalyse und -erkennung.

  Die Bilderkennungsgruppe profitiert von der engen Zusammenarbeit mit der Spracherkennungsgruppe (z.B. in der Erkennung von Bewegungen unter Zuhilfenahme der HMM Technologie oder in der multimodalen Erkennung unter Benutzung der Multi-Stream Verarbeitung) sowie der Gruppe Automatisches Lernen (z.B. in Klassifikation oder Fusion von Klassifikatoren).

  Diese Forschungsresultate haben zu verschiedenen Errungenschaften in neuen Technologien und Anwendungen geführt. Eine Technik der Objektsuche wurde auf das Problem der Suche nach Gesichtern in Bildern und Video angewandt und in einem System zur Echtzeitgesichtssuche implementiert, was z.B. Anwendungen in der automatischen Überwachung oder Indizierung findet. Des weiteren wurde ein Algorithmus zur Lippenverfolgung entwickelt und, in Kombination mit Methoden der Bewegungserkennung, auf visuelle Spracherkennung und Personenauthentifizierung angewandt. Dieser Ansatz wurde mit akustischen Sprachanalysemethoden kombiniert, was zur Realisierung von Systemen zur audio-visuellen Spracherkennung und audio-visuellen Personenauthentifizierung geführt hat, die robuster sind als monomodale Systeme. Mehrere Methoden der Sensorfusion wurden untersucht und angewandt, um monomodale Systeme der Personenverifizierung zu multimodalen Systemen zu kombinieren. Unsere Methode zur Perso-

nenauthentifizierung wurde von der Cerberus AG und der Ibermatica SA in potentielle Prototypen integriert. Desweiteren hat unsere Tätigkeit auf dem Gebiete der Röntgenbildanalyse zu einem System zur Extraktion von artikulatorischen Merkmalen in Röntgenbildsequenzen geführt. Die Gruppe hat ausserdem mehrere Systeme zur Dokumentanalyse untersucht, die es erlauben, Druckschrift, Handschrift und Schreibschrift zu erkennen.

- **Automatisches Lernen, einschliesslich Musterklassifizierung, Datenanalyse und Wissensextraktion.**

  Dieses Gebiet umfasst das Beherrschen aller dem Stand der Technik entsprechenden automatischen Lernmethoden, die in der Künstlichen Intelligenz angewandt werden, mit dem Ziel, konkrete Probleme der Klassifizierung, Mustererkennung und der Fusion von Klassifikatoren zu lösen.

  Die Gruppe verfügt über grundlegende Sachkenntnisse in unterschiedlichen Techniken wie künstlichen neuronalen Netzwerken, Bayesian Netzwerken (statistisches Lernen), Entscheidungsbäumen (Symbolisches Lernen), Support Vector Machinen (Optimierung) und Logischer Analyse von Daten (Boolesche Funktionstheorie).

  Die Synergie zwischen dieser weiten Basis an Lerntechniken und spezifischen Applikationen in der Sprachverarbeitung und der Mustererkennung hat bereits zu einigen originellen Lösungsansätzen und unerwarteten Resultaten geführt. Es muss hingegen ein ziemlicher Forschungsaufwand geleistet werden, um allgemeine Methoden für spezifische Probleme (z.B. mit grossen verrauschten Sprachdatenbanken), anzupassen. Im Besonderen werden erhebliche Forschungsbemühungen für die Zerlegung von grossen Problemen in Gruppen einfacherer Teilprobleme unternommen.

  Gleichermassen werden andere aussichtsreiche Aktivitäten durchgeführt, in denen unser Knowhow in automatischem Lernen, angewandt auf die Vorhersage von Zeitfolgen und dem Design von Systemen zur assistierenden Diagnose, zum Tragen kommt.

- **Systemmanagement, einschliesslich Datenbankenmanagement und Prototypenentwicklung.**

  Die drei oben genannten Forschungsgruppen werden von einer leistungsfähigen Gruppe für Systemmanagement unterstützt, die für das Management der Datenbanken und der Prototypenentwicklung verantwortlich ist. Im Rahmen anwendungsbezogener Projekte arbeitet sie eng zusammen mit den anderen Forschungsgruppen.

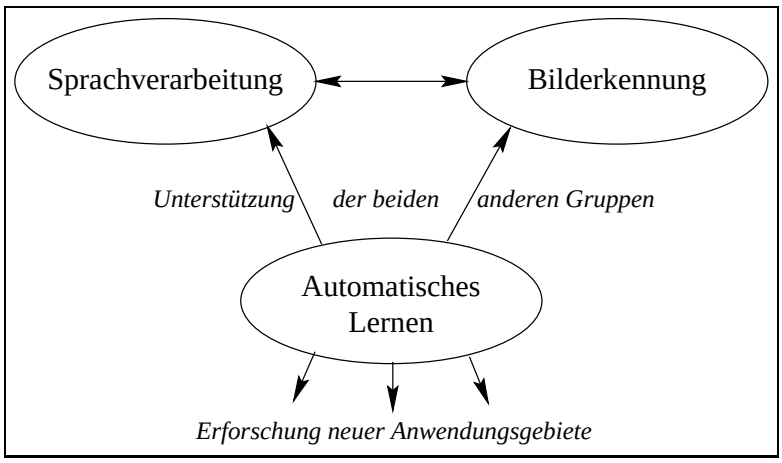


Abbildung 6: *Synergie zwischen den drei Forschungsgruppen.*

Wie weiter oben kurz beschrieben und illustriert in Abbildung 6, wurden die drei Forschungsgruppen definiert, um sich so gut wie möglich zu ergänzen und gleichzeitig eine aktive Zusammenarbeit in verschiedenen Forschungsthemen zu fördern. Sprachverarbeitung und Bilderkennung können sich nicht nur oft ergänzen (z.B. in multimodalen Applikationen), sondern basieren auch vielfach auf gemeinsamen Theorien und mathematischen Werkzeugen und können deshalb von der gegenseitigen Wechselbeziehung profitieren. So stützen sich zum Beispiel einige neue Entwicklungen in der Handschriftenerkennung auf Hidden Markovmodelle, die ursprünglich für die Spracherkennung entwickelt wurden. In ähnlicher Weise werden einige jüngste Entwicklungen in der Multi-Stream Verarbeitung in beiden Gruppen ausgenützt, was auch der Entwicklung von multimodalen Systemen zugute kommt (wie kürzlich am IDIAP anhand einiger Arbeiten über audio-visuelle Spracherkennung gezeigt wurde). Die Gruppe Automatisches Lernen untersucht und entwickelt neue Technologien die gemeinsam von der Sprachverarbeitung, der Bilderkennung und der multimodalen Verarbeitung sinnvoll genutzt werden können. Sie nimmt deshalb die Aufgabe der technischen Unterstützung für die mehr anwendungsorientierten Gruppen wahr. Zum Beispiel wurden neue Methoden, die für die Zerlegung eines Lernproblems in kleinere Teilprobleme vorgeschlagen worden waren, erfolgreich in der Sprecherverifizierung angewandt. Forschung, die zum Ziel hat, Teilmodule von solchen Zerlegungen neu zu kombinieren, trägt direkt zum Problem der Fusion multimodaler Klassifikatoren bei. Weiter ist die Gruppe verantwortlich, neue Anwendungsgebiete ausfindig zu machen, die direkten Nutzen der verfügbaren Technologie erbringen könnten und die wichtig werden könnten für zukünftige Aktivitäten am IDIAP (z.B. Voraussagen von Zeitfolgen). Im letzteren Fall wird Forschung vielfach in Zusammenarbeit mit anderen Instituten ausgeführt, die mehr Erfahrung auf den jeweiligen spezifischen Gebieten haben.

1.3 Teilnahme an Nationalen und Europäischen Forschungsprojekten

In den letzten Jahren hat sich IDIAP besonders intensiv an nationalen Projekten und an Projekten der europäischen Union beteiligt. Das Institut hat oft eine führende Rolle in der Konzeption und beim Koordinieren dieser Projekte gespielt.

Im Jahre 1998, mit Bezug auf die Projekte, die von der europäischen Union im Rahmen des Vierten Programmes für Forschung und Technologie finanziert wurden, war IDIAP aktiv an verschiedenen EU Projekten beteiligt. Dies waren unter anderem die folgenden Projekte:

- Zwei Telematik Projekte:
  - *SpeechDat*, um grosse mehrsprachige Sprachdatenbanken zu standardisieren, zu evaluieren und zu produzieren.
  - *PIoneering Caller Authentication for Secure Service Operation (PICASSO)*, Benutzung von automatischer Sprecherverifizierung für Applikationen im Bank- und Telefonbereich.
- Zwei ESPRIT Langzeitforschungsprojekte:
  - *THematic Indexing of Spoken Language (THISL)*, automatische Indizierung und Zugriff via Sprache auf archivierte Rundfunknachrichten (von BBC).
  - *REcognition of Speech by Partial Information TEchniques (RESPITE)*, Entwicklung von neuen Spracherkennungstechniken die robust sind gegenüber unerwarteter Störungen durch Rauschen und Korruption, um das Einsetzen dieser Techniken in schwierigen Anwendungsgebieten zu ermöglichen (Mobiltelefone und Systeme im Kraftfahrzeug). Dieses Projekt wurde für das Jahr 1998 gewährt, aber offiziell erst am 1. Januar 1999 gestartet.
- Ein TMR Projekt: *Training and Mobility of Researchers*, das zum Ziel hat, die internationale Mobilität der Forscher zu fördern. IDIAP ist einer der Hauptpartner im Rahmen des SPHEAR (SPeech, HEAring and Recognition) Projekts, das ein besseres Verstehen der Verarbeitung des Hörvorgangs zum Ziel hat, um dieses Verständnis zur Verbesserung der automatischen Spracherkennung in ungünstigen Konditionen anzuwenden.

- Zwei COST Projekte: COST249 für automatische Spracherkennung übers Telefon, und COST250, für automatische Sprecherverifizierung übers Telefon. Es ist hier wichtig zu vermerken, dass IDIAP im Vergleich zu nichtschweizerischen Partnern eine substantielle Unterstützung vom BBW im Rahmen der zwei COST Projekte erhalten hat, was uns erlaubte, unseren Sachverstand in Sprach- und Sprechererkennung zu vertiefen und somit neue industrielle und europäische Projekte zu initiieren.

- Ein ACTS (Advanced Communication Technologies and Services) Projekt über multimodale Personenverifizierung für Anwendungen im Teleservice- und Sicherheitsbereich (M2VTS).

- Ein Projekt innerhalb des SOCRATES/ERASMUS Programmes, in dem IDIAP als einzige Institution die Schweiz vertritt. Das Ziel dieses Projekts ist ein europäisches Aufbaustudium (Master Course) in Sprachverarbeitung zu definieren und zu initiieren. Gemeinsame Kernkurse werden in allen repräsentativen europäischen Ländern abgehalten werden, gefolgt von Vertiefungskursen und Projekten die in ausgehählten Instituten erfolgen werden.

Im Rahmen nationaler Projekte war IDIAP vorwiegend in verschiedenen Projekten des Nationalfonds tätig, die hauptsächlich für die Ausbildung von Doktoranden dienen, unter anderem:

- AV-COM: *Audio-Visual Combination.* Audiovisuelle Kombinierung und multimodale Erkennung (gewährt im Jahre 1997 für Ausrüstung).

- MULTICHAN: *Non-stationary multichannel signal processing.* Nichtstationäre Mehrkanal Signalverarbeitung zur Spracherkennung.

- SV-UCP: *Speaker Verification based on User-Customized Password.* Sprecherverifizierung basierend auf benutzerangepasstem Passwort (im Jahre 1998 gewährt) .

- INSPECT: *INtegrating SPeech (acoustic and linguistic) ConsTraints for enhanced recognition systems.* Integrierung akustischer und sprachlicher Nebenbedingungen für verbesserte Erkennungssysteme (im Jahre 1998 gewährt), in Zusammenarbeit mit der EPFL (Dr. Martin Rajman, DI/LIA).

- BN-ASR: *Modelling the hidden dynamic structure of speech production in a unified framework for robust automatic speech recognition.* Modellierung der verborgenen dynamischen Struktur der Spracherzeugung zur robusten automatischen Spracherkennung (im Jahre 1998 gewährt).

- ARTIST: *Articulatory Representation Towards Improved Speech Technology.* Untersuchung und Benutzung von Artikulationsmerkmalen in Spracherkennungssystemen.

- FaceX: *Facial Expression Recognition through Temporal and Appearance Based Models.* Automatische Erkennung des Gesichtsausdrucks mittels räumlich-zeitlicher visueller Modellierung (im Jahre 1998 gewährt).

- GLAD: *Generalization of "Logical Analysis of Data" techniques.* Generalisierung der "Logical Analysis of Data" Techniken, die zur Klassifizierung angewendet werden.

- *Compact hardware-friendly neural networks.* Kompakte Hardware-freundliche neuronale Netzwerke. Der jetzige Forschungsschwerpunkt befasst sich mit der Fusion von Klassifikatoren.

- ZEPHYR: *Time series prediction with hybrid Markov models.* Voraussage von Zeitfolgen mittels hybriden Markovmodellen, z.B. finanzielle Vorhersagen, Vorhersagen über Lawinengefahr.

- CARTANN: *Cartography by Artificial neural networks.* Kartographie mittels neuronaler Netzwerke (im Jahre 1998 gewährt), in Zusammenarbeit mit der Universität Lausanne (Prof. Michel Maignan, Departement der Erdwissenschaften).

Im Jahre 1998 wurde VOXCom, eine direkte Spin-Off-Firma von IDIAP, gegründet. Im selben Jahr wurde InfoVOX, ein neues CTI (Comission for Technology and Innovation) Projekt mit den Partnern IDIAP, EPFL, Swisscom, VOXCom S.A. und Omedia S.A. akzeptiert, wobei der Starttermin auf den 1. März 1999 festgelegt wurde. Eines der Ziele von InfoVOX ist, Forschung und Entwicklung auf dem Gebiet interaktiver Sprach-Server zu betreiben. Diese kommen in Anrufzentren und Applikationen in der Computertelefonie zur Anwendung. Die allgemeine Aufgabe beinhaltet die Entwicklung von interaktiven Systeme mit Sprachabfrage (Interactive Voice Response Systems, IVR), um auf grosse Informationsdatenbanken zugreifen zu können. Im aktuellen Projekt beabsichtigen wir, uns hauptsächlich auf Internetdatenbanken zu konzentrieren, wobei wir als Fallstudie die Entwicklung einer Sprachschnittstelle für die WWW-Seite des Fremdenverkehrsbüros in Martigny (http://www.martigny.ch) in Angriff nehmen.

1.4 Zusammenarbeit mit anderen Institutionen und Unternehmen

In den letzten Jahren hat IDIAP enge Kontakte mit Forschungsorganisationen, Universitäten und der Industrie unterhalten, die in denselben Forschungs- und Entwicklunggebieten tätig sind. Diese Kontakte kamen vielfach in Folge eines gemeinsamen und mit Erfolg abgeschlossenen Projektes zustande oder basieren auf persönlichen Kontakten, die durch regelmässige Austausche gefestigt wurden. Wir möchten an dieser Stelle einige Beispiele erwähnen:

- Die aktive Zusammenarbeit zwischen IDIAP und Swisscom (eine kurze Beschreibung mit Ergebnissen folgt später).
- Enge Partnerschaften mit akademischen Institutionen wie der EPFL (an der Hervé Bourlard auch Professor ist) und der Universität Genf. Momentan laufen mehrere Forschungprojekte in enger Zusammenarbeit mit der EPFL. Mehrere Doktoranden am IDIAP sind oder werden mit der EPFL affiliiert sein. Wir initiieren zur Zeit auch neue Projekte mit der Universität Genf.
- IDIAP hat jetzt sehr gute Kontakte mit mehreren Unternehmen, die durch europäische Projekte zustande kamen. Unter anderem mit Cerberus (CH), BBC (UK), DaimlerChrysler (D), Thomson (F), MatraNortel (F), Ibermatica (E) und mehreren anderen Universitäten, einschliesslich der Universität Cambridge, der Universität Sheffield, der Faculté Polytechnique de Mons (BE), der Universität Surrey und dem IMT (Neuenburg).
- Wir möchten an dieser Stelle die Universität Rutgers (RUTCOR) und das International Computer Science Institute (ICSI) erwähnen, zu denen dank persönlicher Kontakte und regelmässigem Informationsaustausch gute Beziehungen aufgebaut wurden.
- Erst kürzlich starteten wir eine auf 4 Jahre ausgelegte Zusammenarbeit, die im Rahmen eines von der Catalyst Foundation (USA) finanzierten Projektes stattfindet. Das Ziel des Projektes ist, zusammen mit der Johns Hopkins Universität (Baltimore, USA) und dem Indian Institute of Technology (Delhi), eine Micropower analoge VLSI Implementierung für ein kontinuierliches Spracherkennungssystem zu implementieren.

1.5 Ausbildung und Regionale Entwicklung

Neben qualitativ hochwertiger Forschung und Entwicklung sehen wir als weitere wichtige Aufgaben:

Ausbildung und Betreuung von Doktoranden (meistens affiliiert mit der EPFL, der Universität Lausanne oder der Universität Genf) und Forschern, als auch Kurzzeit- oder Langzeitbesuchern von akademischen Institutionen (einschliesslich Fachhochschulen) und der Industrie. Zum Beispiel ist IDIAP (als einzige Institution die Schweiz) zusammen mit anderen europäischen Partnern (im Rahmen des europäischen Projektes SOCRATES/ERASMUS) an der Ausarbeit des Inhaltes eines europäischen Master-Kurses in Sprachtechnologie beteiligt (Kernkurse könnten in allen

Ländern abgehalten werden, gefolgt von weiterführenden Kursen und Projekten die in speziellen Ländern stattfinden würden). Die ersten Testläufe dieses Master-Kurses sollten im Jahre 1999 beginnen.

Zur Zeit sind am IDIAP 13 Doktoranden, als auch mehrere Studenten, die ihre Diplomarbeit vorbereiten, beschäftigt. Letztere studieren an der EPFL, EURECOM (F), ENST (F) oder der ETS (Ecole Technique Supérieur) in Sion. Jedes Jahr steht IDIAP ein Budget für den Aufenthalt externer Forscher oder Studenten zur Verfügung, über einen Zeitraum von 36 Monaten (12 Monate für jede Gruppe).

Technologietransfer und Unterstützung der Industrie mit folgender Motivierung: (1) Damit sich die Industrie auf dem neusten Stand der Technik halten kann (da es sogar für grosse Unternehmen oft zu teuer ist, betriebsinterne Kompetenz in allen wichtigen technologischen Gebieten beizubehalten), und (2) das Entwickeln und Testen von Prototypen für Applikationen, die von besonderem Interesse einiger unserer Sponsoren sind (z.B. Swisscom). Die Strategie ermöglicht uns, gleichermassen qualitative und quantitative Analysen durchzuführen, die Anwendbarkeit von gegenwärtigen Basistechnologien zu verbessern, als auch Technologien in Pilotsysteme zu integrieren und sich für einen aktiven Technologietransfer einzusetzen. Es ist das Ziel von IDIAP, durch offene und intensive Zusammenarbeit mit der Industrie, eine wichtige Rolle im Voranbringen der wirtschaftlichen Entwicklung des Kanton Wallis zu spielen.

In diesem Rahmen wurde im Juli 1998 VOXCom S.A. als direkte Spin-Off-Firma von IDIAP mit Beteiligung der Stadt Martigny lanciert. Der Schwerpunkt von VOXCom wird die Entwicklung und Anwendung von Softwareprodukten und Servicen sein, um den speziellen Bedürfnissen von Unternehmen, die Computertelefonie, Anrufzentren (Callcenters) und Integrationssupport benötigen, Rechnung zu tragen. VOXCom's Hauptaufgabe ist die Integration und Umsetzung von am IDIAP entwickelter Basistechnologie in Lösungsansätze für Informationssysteme und andere Anwendungen im Umfeld von Anrufzentren.

IDIAP ist darüberhinaus regelmässig an der Organisierung wissenschaftlicher Veranstaltungen beteiligt, z.B.

- Der ersten internationalen Konferenz über "Audio- and Video-based Biometric Person Authentication" (AVBPA), die vom 12-14, März 1997 in Crans-Montana stattfand.
- "Journées d'Etude sur la Parole" (JEP'98), die in Martigny vom 15-19 Juni 1998 abgehalten wurde.

1.6 Veröffentlichungen

Die Leistung eines Forschungsinstituts wird hauptsächlich gemessen an der Anzahl, vor allem aber auch der Qualität, der Veröffentlichungen. In den Jahren 1997 und 1998 wurden von IDIAP die folgenden Veröffentlichungen geschrieben (für mehr Details, siehe letztes Kapitel dieses Berichtes):

- 2 Bücher (in Vorbereitung)
- 7 Kapitel in Büchern
- 10 Artikel in Fachzeitschriften
- 45 Artikel in internationalen Konferenzen
- 28 interne Forschungsberichte

1 General Overview of the Institute

1.1 Introduction

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

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

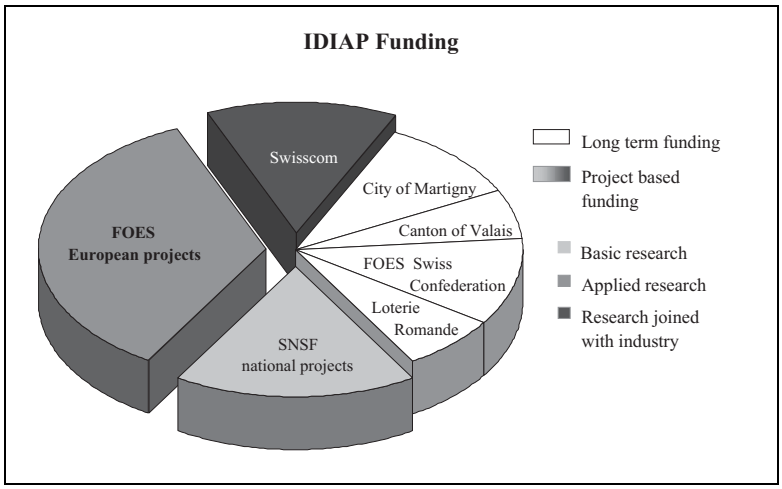


Abbildung 7: *Relative distribution of IDIAP funding in 1998.*

Today, IDIAP is primarily funded by long-term support from the the Swiss Confederation (federal Office for Education and Science, FOES), the State of Valais, the City of Martigny, and Swisscom. The "Loterie Romande" is also supporting our research efforts with annual grants. In addition, IDIAP receives substantial research grants from the Swiss National Science Foundation (SNSF) for basic research projects (mainly for PhD students), and from FOES in the framework of European projects. The relative distribution of IDIAP's funding in 1998 is illustrated in Figure 7.

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

The general management structure of IDIAP is illustrated in Figure 8 and is composed of a Foundation Council, a Board of Directors, and a Scientific Committee (advising the Board of Directors). It is also intended to set up an Economic Relations Committee, which will be responsible for publicizing IDIAP's research results across the industrial world, as well as providing IDIAP with new research opportunities of particular interest to industry.

The activities carried out at IDIAP can be described as follows: research and development activities, participation in European and national research projects, collaborations with organisations and

companies, and teaching and training activities. IDIAP's mission therefore consists in:

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

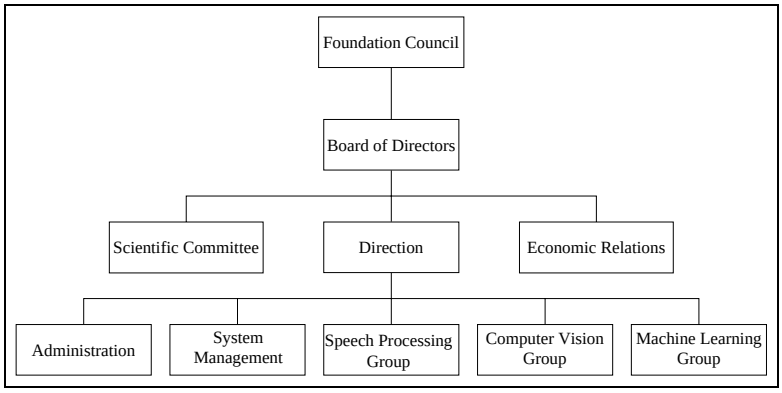


Abbildung 8: *IDIAP Structure.*

As further discussed below, 1998 has been a good year for our Institute, and IDIAP's activities have continued to flourish: the number of national and international projects, as well as partnerships with academic institutions, have grown. Moreover, thanks to the continued support of our authorities, and to our most competent personnel, motivated to the highest level, IDIAP is recognized as a highly sought partner in the well defined areas (speech processing, computer vision, machine learning) they decided to focus on. It is now our job to continue to concentrate our research and development activities on those areas, while fostering technology transfer through industrial partnerships (including VOXCom, a spin-off of IDIAP initiated in July'98 – see later).

1.2 Research and Development Activities

The main 1998 research activities of IDIAP, which focus on medium-long term objectives, are described in detail in the present Activity Report. Focusing on a few, well defined, research axes, along the general theme of **multimodal interaction**, IDIAP carries out fundamental research and develops prototype systems (to validate its research results) along three complementary directions:

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

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

recognition systems, involving complex pronunciation dictionaries and rules, as well as advanced grammatical constraints, are also developed and tested.

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

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

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

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

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

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

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

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

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

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

  The main goal of this group is to maintain a strong expertise in advanced techniques that have been identified as being of direct interest to current and future work at IDIAP. This involves

research in several areas as diverse as Bayesian learning, artificial neural networks, decision trees, support vector machines, and logical analysis of data.

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

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

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

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

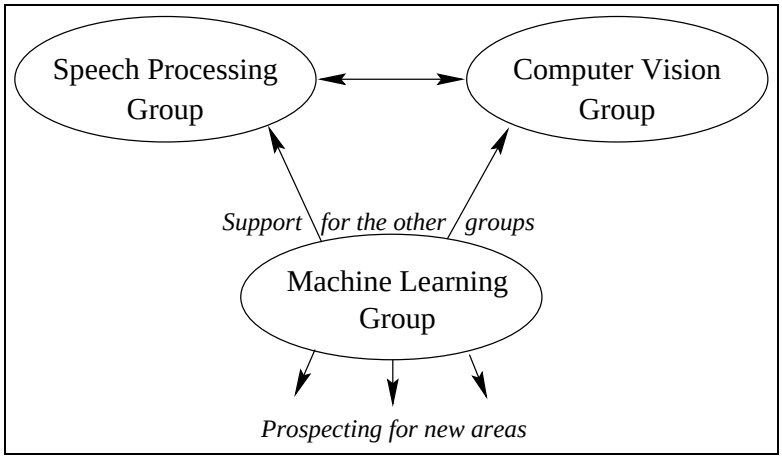


Abbildung 9: *Inter-dependencies between the three research groups.*

As briefly described above, and as illustrated by Figure 9, the activities in the three research groups have been defined to be as complementary as possible, while fostering active collaboration across the different research themes.

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

The Machine Learning group provides additional theoretical support to the two other (more application oriented) groups by investigating new technologies that are common and useful to speech processing, computer vision, and multimodal processing. For example, in 1998, the new methods

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

1.3 Participation in National and European Community Research Projects

The activity of IDIAP within the framework of national and European Community projects has been particularly intense for the last few years, and IDIAP has played a leading role in the conception and coordination of many projects.

In 1998, with regards to projects funded by the European Community (and by FOES for the Swiss partners) in the framework of the fourth Program for Research and Technology, IDIAP was actively involved in several projects, including:

- Two Telematics project:
  - SpeechDat to produce, standardize, and evaluate large multi-lingual speech corpora. In particular, this project allowed IDIAP to develop speech databases for French and German as spoken in Switzerland.
  - PIoneering Caller Authentication for Secure Service Operation (PICASSO) on the use of automatic speaker verification systems in calling card and personal communication assistant applications.
- Two ESPRIT Long Term Research projects:
  - THematic Indexing of Spoken Language (THISL) on automatic indexing and vocal access of recorded broadcast news (from BBC).
  - REcognition of Speech by Partial Information TEchniques (RESPITE) on the development of novel speech recognition techniques that are truly robust to unanticipated noise and corruption, and to deploy these techniques in two application areas (cellular phones and in-car systems). This project was granted in 1998 but officially started on January 1, 1999.
- In the TMR (Training and Mobility of Researchers) program that aims to promote international mobility of researchers, IDIAP is one of the key partners of the SPHEAR (SPeech, HEAring and Recognition) project which has the objective of a better understanding of auditory processing, and to deploy this understanding in automatic speech recognition in adverse conditions.
- Two COST (European Cooperation in the field of Scientific and Technical Research) projects: COST249 on automatic speech recognition over the telephone, and COST250 on automatic speaker verification over the telephone. In comparison to non-Swiss partners, it is important to note here that IDIAP has received substantial funding from FOES in the framework of these two COST projects, which allowed us to significantly boost our expertise in speech and speaker recognition, and to initiate many new industrial and European contacts.
- One ACTS (Advanced Communications Technologies and Services) project on multimodal verification for teleservices and security applications (M2VTS).
- One within the SOCRATES/ERASMUS program, in which IDIAP is the only Swiss representative. The goal of this project is to define and initiate a European Masters program in language and speech. Common core courses will be given in all the represented European countries, followed by specialised courses and projects which will be given in specific institutions.

In the framework of national projects, IDIAP was mainly involved in several Swiss National Science Foundation projects (mainly for the education and training of PhD students), such as:

- AV-COM: Audio-visual combination (granted in 1997, for equipment)
- MULTICHAN: Non-stationary multichannel signal processing (granted in 1997)
- SV-UCP: Speaker Verification based on User-Customized Password (granted in 1998)
- INSPECT: INtegrating SPeech (acoustic and linguistic) ConsTraints for enhanced recognition systems (granted in 1998), in collaboration with EPFL (Dr. Martin Rajman, DI/LIA)
- BN-ASR: Modelling the hidden dynamic structure of speech production in a unified framework for robust automatic speech recognition (granted in 1998)
- ARTIST: Articulatory representation towards improved speech technology
- FaceX: Facial Expression Recognition through Temporal and Appearance Based Models (granted in 1998)
- GLAD: Generalization of "Logical Analysis of Data" techniques
- Compact hardware-friendly neural networks; now mainly focusing on mixture of experts
- ZEPHYR: Time series prediction with hybrid Markov models
- CARTANN: Cartography by Artificial neural networks (granted in 1998), in collaboration with University of Lausanne (Prof. Michel Maignan, Earth Sciences Department).

Finally, although IDIAP used to be involved in a CTI (Commission for Technology and Innovation) project, 1998 was a transition year during which VOXCom, a direct spin-off of IDIAP, was initiated. Around VOXCom, a new CTI project (InfoVOX, involving IDIAP, EPFL, Swisscom, VOXCom S.A. and Omedia S.A.) was defined, submitted, and granted by CTI with a start date of March 1, 1999. One of the goals of InfoVOX is to do further research and development in the field of interactive voice servers, with applications in the key area of call centres for computer telephony applications. The generic application is the development of Interactive Voice Response (IVR) systems (interactive vocal query systems) to access large information databases. In the current project, we intend to mainly focus on internet databases, and as a testbed to develop a voice interface to the web page available from the tourist bureau of Martigny (`http://www.martigny.ch`).

Most of these projects are discussed in more detail in the present report.

1.4 Collaboration with other Organisations and Companies

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

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

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

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

- More recently, in the framework of a project funded by the Catalyst Foundation (USA), we started a 4 years collaboration with Johns Hopkins University (Baltimore, USA) and the Indian Institute of Technology (Delhi, India) on micropower analog VLSI implementation of continuous speech recognition systems.

1.5 Training Activities and Regional Development

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

Training and supervision of PhD students (most of the time affiliated with EPFL, University of Geneva, or University of Lausanne) and postdoctoral fellows, as well as short-term or medium-term visitors from academia (including ETS) and industry. As an example of this specific concern, IDIAP is currently working (as the only Swiss representative) with other European partners (in the framework of a European Socrates/Erasmus project) on defining the content of a European Masters in Language Technology (in which core courses would be given in all countries, followed by specialised courses and projects that would be taken in pre-defined countries). The first test of this Masters program should start in 1999.

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

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

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

Finally, IDIAP is regularly involved in the organisation of scientific events such as:

- The first international conference on "Audio-and Video-based Biometric Person Authentication" (AVBPA), held in Crans-Montana on March 12-14, 1997.

- "Journées d'Etude sur la Parole" (JEP'98), held in Martigny on June 15-19, 1998.

2 Staff

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

Phone: +41 – 27 – 721 77 11

Fax: +41 – 27 – 721 77 12

Internet: http://www.idiap.ch

2.1 Scientific Staff

Persons at IDIAP in 1998 or as of this writing:

Mr.	Johan Myhre Andersen Johan.Myhre.Andersen@idiap.ch	research assistant until October 98
Dr.	Souheil Ben Yacoub Souheil.Ben-Yacoub@idiap.ch	research scientist +41 – 27 – 721 77 38
Ms.	Giulia Bernardis Giulia.Bernardis@idiap.ch	research assistant +41 – 27 – 721 77 36
Mr.	Olivier Bornet Olivier.Bornet@idiap.ch	System Management group leader +41 – 27 – 721 77 40
Dr.	Hervé Bourlard Herve.Bourlard@idiap.ch	Professor EPFL, Director +41 – 27 – 721 77 20
Mr.	Gilles Caloz Gilles.Caloz@idiap.ch	research assistant until June 98
Mr.	Thierry Collado Thierry.Collado@idiap.ch	development engineer +41 – 27 – 721 77 42
Mr.	Beat Fasel Beat.Fasel@idiap.ch	research assistant +41 – 27 – 721 77 23
Mr.	Frank Formaz Frank.Formaz@idiap.ch	development engineer +41 – 27 – 721 77 28
Mr.	Dominique Genoud Dominique.Genoud@idiap.ch	research assistant +41 – 27 – 721 77 26
Mr.	Nicolas Gilardi Nicolas.Gilardi@idiap.ch	research assistant +41 – 27 – 721 77 47
Mr.	Hervé Glotin Herve.Glotin@idiap.ch	research assistant +41 – 27 – 721 77 33
Mr.	Frédéric Gobry Frederic.Gobry@idiap.ch	research assistant +41 – 27 – 721 77 31

Ms.	Astrid Hagen Astrid.Hagen@idiap.ch	research assistant +41 – 27 – 721 77 34
Ms.	Katrin Keller Katrin.Keller@idiap.ch	research assistant +41 – 27 – 721 77 37
Mr.	Christopher Kermorvant Christopher.Kermorvant@idiap.ch	research assistant +41 – 27 – 721 77 46
Mr.	Sacha Krstulović Sacha.Krstulovic@idiap.ch	research assistant +41 – 27 – 721 77 43
Dr.	Mikko Kurimo Mikko.Kurimo@idiap.ch	research scientist +41 – 27 – 721 77 41
Mr.	Bertrand Liardon Bertrand.Liardon@idiap.ch	development engineer +41 – 27 – 721 77 48
Dr.	Jürgen Lüttin Juergen.Luettin@idiap.ch	Computer Vision group leader +41 – 27 – 721 77 27
Dr.	Djamila Mahmoudi Djamila.Mahmoudi@idiap.ch	research scientist +41 – 27 – 721 77 24
Dr.	Gilbert Maître Gilbert.Maitre@idiap.ch	research scientist until July 98
Mr.	Johnny Mariéthoz Johnny.Mariethoz@idiap.ch	development engineer +41 – 27 – 721 77 44
Dr.	Eddy Mayoraz Eddy.Mayoraz@idiap.ch	Machine Learning group leader +41 – 27 – 721 77 29
Mr.	Perry Moerland Perry.Moerland@idiap.ch	research assistant +41 – 27 – 721 77 32
Dr.	Chafic Mokbel Chafic.Mokbel@idiap.ch	Speech Processing group leader +41 – 27 – 721 77 30
Dr.	Houda Mokbel Houda.Mokbel@idiap.ch	research scientist +41 – 27 – 721 77 51
Mr.	Miguel Moreira Miguel.Moreira@idiap.ch	research assistant +41 – 27 – 721 77 45
Dr.	Andrew Morris Andrew.Morris@idiap.ch	research scientist +41 – 27 – 721 77 35
Mr.	Bojan Nedić Bojan.Nedic@idiap.ch	research assistant +41 – 27 – 721 77 25
Mr.	Todd Stephenson Todd.Stephenson@idiap.ch	research assistant +41 – 27 – 721 77 52
Dr.	Georg Thimm Georg.Thimm@idiap.ch	research scientist +41 – 27 – 721 77 39

2.2 Visitors

Prof.	Ethem Alpaydin Ethem.Alpaydin@idiap.ch	Bogaziçi University, Istanbul, Turkey from February to June 1998
Ms.	Heidi Christensen Heidi.Christensen@idiap.ch	University of Aalborg, Denmark +41 – 27 – 721 77 50
Dr.	Sergio Curinga Sergio.Curinga@idiap.ch	University of Genova, Italy from September 97 to January 98
Mr.	Dan Gildea Dan.Gildea@idiap.ch	ICSI, Berkeley, USA from June 98 to July 98
Prof.	Mikhael Kanevski Mikhael.Kanevski@idiap.ch	IBRAE, Moscow, Russia +41 – 27 – 721 77 49
Dr.	Ascension Vizinho Ascension.Vizinho@idiap.ch	University of Sheffield, UK from April 98 to September 98

2.3 Students

Ms.	Ana Merchan	Ana.Merchan@idiap.ch	from October 97 to April 98
Mr.	Samuel Vannay	Samuel.Vannay@idiap.ch	July 97 to January 98

2.4 Administrative Staff

Mr.	Yann Briguet Yann.Briguet@idiap.ch	annotator until August 98
Ms.	Sylvie Millius Sylvie.Millius@idiap.ch	secretary +41 – 27 – 721 77 21
Ms.	Nadine Rousseau Nadine.Rousseau@idiap.ch	secretary +41 – 27 – 721 77 22

3 Research Activities

3.1 Speech Processing Group

Group Leader: Chafic Mokbel

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

3.1.1 Overview of the Speech Processing group activities

As illustrated in Figure 10, the activities of the Speech Processing group expand along two main axes: on one hand, the different research themes and underlying technology, and on the other hand, the development levels ranging from fundamental research to prototype development. The goal of prototypes is to demonstrate the added value of technology to different services, while providing important feedback to research. Nevertheless, the main focus of the Speech Processing group is kept at the technological level.

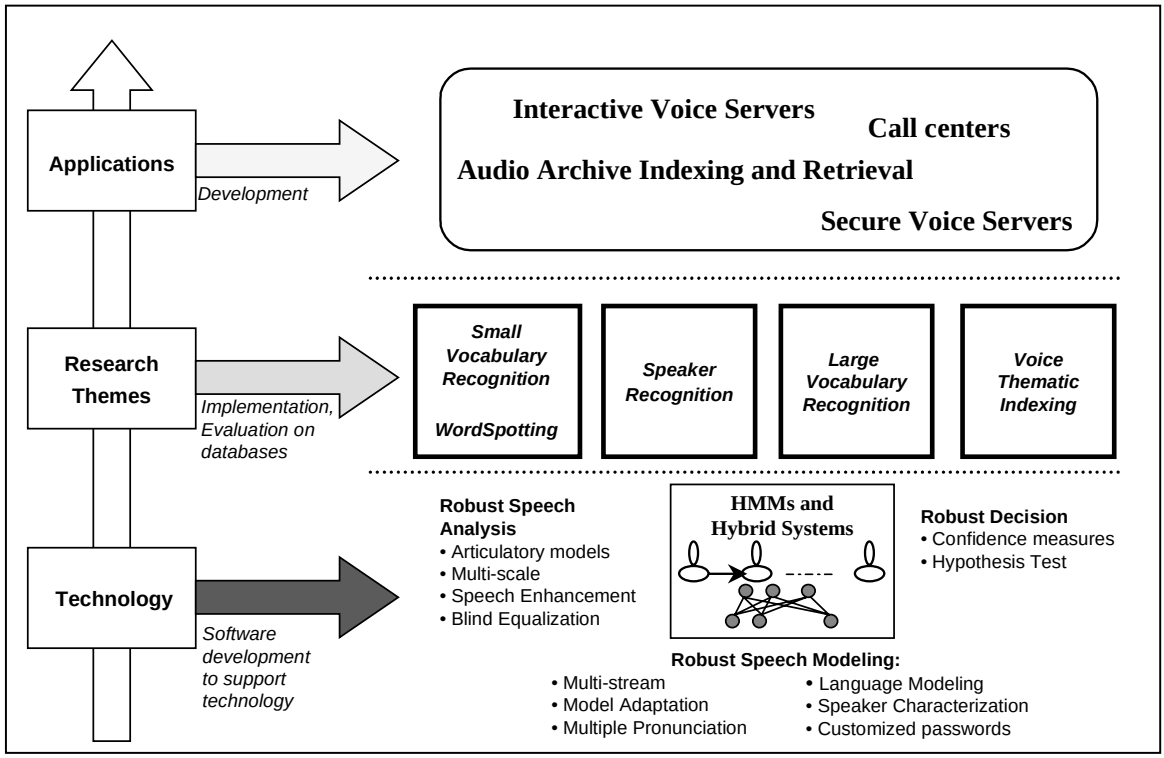


Abbildung 10: *Main activities within the speech group.*

Technological level

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

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

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

Research themes

As shown in Figure 10, four main research themes, corresponding to the main application domains, have been defined and were investigated at IDIAP in 1998:

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

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

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

- **Voice Thematic Indexing**, which has been an important research focus at IDIAP in 1998.

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

Applications and development

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

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

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

In the following, we will start by briefly describing the basic speech recognition technology which is the basis for IDIAP's research. This technology represents the state-of-the-art. It is used in many other laboratories and industries, and is at the basis of most of the current products. Then, the different research themes are presented in more detail, followed by a short description of some related ongoing projects of year 1998. As a conclusion, the main achievements and results obtained in 1998 will be discussed.

3.1.2 Base Technology Tools

A typical speech recognition system is illustrated in Figure 11. At its input, the speech signal obtained at the microphone output represents the amplitude of the waveform as a function of time. After digitization (typically at 8kHz for telephone speech), this signal is analyzed to produce a sequence of feature vectors defining an information measure over time.

Analysis module

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

Acoustic modeling layer

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

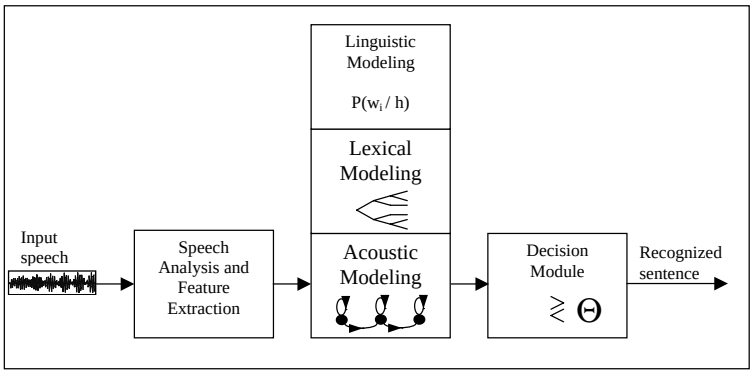


Abbildung 11: *Typical speech recognition system.*

physical model is available for the recognition of speech signal, stochastic models (sometimes referred to as "ignorance-based models") are generally used. In this case, it is assumed that speech sequences result from a Markovian process. Hence, Hidden Markov Models (HMMs) are popularly used to model speech units. Figure 12 represents a typical HMM, modeling a time-information distribution. In this framework, the speech signal is described as the output of a stochastic finite-state automaton built up from a discrete set of states. The system changes its internal state following a discrete probability density function known as transition probabilities. These are defined as the set of probabilities associated with the possible transitions of the underlying Markovian automaton. But the internal states of the underlying Markov process are not directly observable. Only the outputs of the process (the acoustic vectors) are observed and, given the frequency variability, these will be assumed to be themselves stochastic functions of the hidden internal states. Consequently, when in a specific state, the HMM system is assumed to emit measurable stationary observations according to a specific output distribution function. These output distributions are usually parameterized in terms of either Gaussian mixtures, or Artificial Neural Networks (ANN, in the case of hybrid HMM/ANN systems). The set of transition and output distribution parameters can be estimated through efficient training algorithms referred to as *Baum-Welch* and *Viterbi* training. These algorithms make use of large sets of training data constituted of acoustic sequences and their associated word sequences.

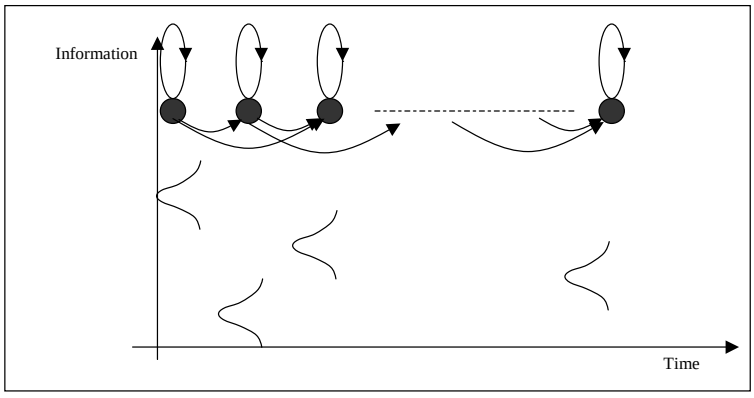


Abbildung 12: *HMM modeling of time-information distribution.*

Lexical modeling layer

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

Language modeling layers

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

Other layers and modules

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

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

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

Speaker recognition/verification

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

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

3.1.3 Small / Medium Vocabulary Robust Speech Recognition

In real-life applications, the speech signal presented at the input of a recognition system is often degraded with additive noise and/or channel distortions (convolutional noise). Moreover, the user may pronounce words that do not belong to the target lexicon, especially in the case of small/medium size lexicons. Robustness to noise and out-of-vocabulary (OOV) words is a key factor to the success of automatic speech recognition in real-life applications. Considering Figure 11, robustness can be improved by acting on several modules of the system. Consequently, the main research directions at IDIAP in 1998 towards robust small/medium vocabulary speech recognition can be classified as follows:

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

  IDIAP has a large expertise with most of these features. In 1998, we assessed the robustness of RASTA-PLP and *J*-RASTA-PLP features to several types of additive noise and at different signal to noise ratios (SNR) on telephone databases. Another feature set was derived from articulatory modeling and was experimented : we developed an algorithm to perform acoustic to articulatory inversion in the case of the DRM model. Another approach that we also currently investigate is the use of multi-resolution features (defining multiple time scale features). These features are then combined directly into a single feature vector for further processing or integrated in the framework of the multi-stream approach described thereafter.

- **Developing preprocessing techniques** to reduce the effects of disturbances in the observed speech signals. Speech enhancement or blind channel effects equalization are used for this purpose. Spectral subtraction, based on the estimation of the noise spectrogram is also a very popular approach to reduce the effects of additive noise. However, since short-term estimation of signal-to-noise ratio (SNR) is not easy, this approach is best suited for stationary noise. In 1998, a variant of spectral subtraction has been developed at IDIAP in order to be combined with more recent developments based on multi-stream or missing data approaches.

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

  In 1998, IDIAP also started looking at blind equalization techniques using adaptive filtering.

- **Developing modeling architectures**, since robustness can be improved by using different **HMM topologies and models**. In 1998, IDIAP was intensively working on **multi-stream** and **missing data** approaches. The basic idea of the multi-stream approach is to represent the speech data in terms of independent subsets of feature vectors and to develop independent models for each subset. Since noise may affect a stream more than another, the robustness is increased by processing the different streams separately and recombining the output of the different models appropriately. A particular case of this approach is the multi-band technique, where the full frequency band is no longer considered as a single feature but as a set of sub-band features. Missing data theory is somehow related to this idea by assuming that noisy features are simply irrelevant. Hence, only the clean features, localized in time and frequency, are used in the calculation of the HMM probabilities. The two main issues underpinning these approaches are (1) how to identify the noisy or missing streams, or to measure their reliability, and (2) how to integrate this information into the decoding process. In the following, multi-stream and

missing data approaches are further detailed, as well as the techniques initially developed to address these two issues.

1. *Multi-Stream Approach*
   As exposed earlier, the acoustic processing module delivers a sequence of acoustic feature vectors that each describes local components of the speech signal. HMMs then assume piecewise stationarity of the signal, and each stationary segment is associated with a specific HMM state. As illustrated in Figure 13, the multi-stream approach avoids this limitation by constructing a model for each subset of features, or feature stream. It thereafter combines the probabilities resulting from the different streams. This combination can be operated at different levels of the modeling process, e.g., HMM states, sub-word units (phonemes or syllables), words, or sentences.

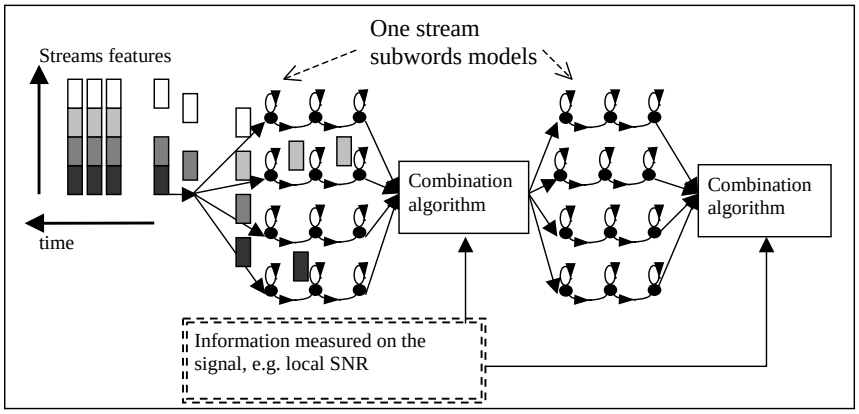


Abbildung 13: *Multi-Stream approach.*

2. *Missing Data Approach*
   The missing data approach is based on the assumption that some of the features are not observed, or are simply not relevant for recognition (too noisy or not carrying any information). For instance, in practice, features are never completely missing but highly disturbed by additive noise. In that case, the likelihoods associated with the current feature vector are estimated on the basis of the remaining features only (e.g., by estimating marginal distributions).

3. *Quantification of Noisy or Missing Streams*
   Localization and quantification of the noisy or missing streams remains a key issue to the success of both multi-stream and missing data, and several related approaches have been investigated at IDIAP in 1998. The first approach is based on sequential estimation of the noise characteristics, based on the estimation of first and second-order moments. To do this sequential estimation, the acoustic frames are classified as speech or noise through a statistical test. A second approach is based on the estimation of the local harmonicity degree. For voiced sounds, the harmonicity degree is defined as the ratio between the maximum R1 of the correlation function within a pitch period and the energy R0 of the signal.

4. *Integrating SNR Information into the Modeling Process*
   The estimated SNR must be integrated into the modeling process. This information will be used in the combination module of the multi-stream approach, or to estimate the set of likelihoods in the missing data approach. The first solution studied in 1998 consists in selecting the streams that are considered clean (not highly disturbed), or combining the

likelihoods of the different streams according to a weighted sum, in which the weights are proportional to the estimated stream-specific SNR. Another solution, referred to as the *full combination approach*, performs the combination as a weighted sum over all possible stream subsets.

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

All the preceding approaches have been studied and developed within the framework of several national and European projects. Some of these developments were also carried out in the framework of our collaboration with SWISSCOM (see below). We will now briefly review some of the most representative projects in the field of small/medium vocabulary speech recognition.

♢ MULTICHAN – Non-stationary MULTI-CHANnel signal processing

Funding: Swiss National Science Foundation (SNSF)

Duration: January 98 – December 99

Persons involved: Katrin Keller and Andrew Morris

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

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

Achievements: A reference speech recognition system based on the HTK software platform was developed. During the tests on the Numbers95 telephone database it proved to be one of the best systems. Furthermore, multiple time scales speech analysis was implemented. Different analysis were combined into feature vectors, which were again transformed to lower dimensional vectors. The resulting performance is very encouraging since this method showed to be robust in presence of noise. Work will be pursued in this direction to select the most informative features from the different scales.

In the case of sub-band ASR, a novel technique for speech recognition in noisy environments, referred to as the "Full Combination" method, was developed to avoid independent sub-band processing. Recognition results for speech in noise were better than for any previous sub-band ASR method. This method also has considerable scope for further improvement through improved methods for local data reliability estimation. Full combination approach has been developed within both MULTICHAN and SPHEAR projects.

♢ SPHEAR – SPeech, HEAring and Recognition

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

Duration: March 1998 –February 2002

Persons involved: Astrid Hagen and Christopher Kermorvant

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

Achievements: The basic multi-stream processing and the missing data approach have been implemented within two different speech recognition systems (HTK and STRUT). The first system (HTK) is based on standard HMM modeling with Gaussian mixture as output distributions while the second system (STRUT) implements a hybrid HMM/ANN system in which the output distributions are approximated by Artificial Neural Networks (ANN, in our case a Multilayer Perceptron, or "MLP"). The basic versions of multi-stream and missing data have shown to achieve competitive performance on Numbers95 disturbed by additive noise extracted from the NOISEX database. Important effort has also been put into the definition of an experimental protocol allowing a fair comparison of the different algorithms. Improved algorithms for feature reliability estimation and integration were proposed, as described in the introduction of this section, resulting in significant improvements with respect to the basic versions of the algorithms. Spectral subtraction and J-RASTA-PLP were evaluated as reference systems. Figure 14 reflects some of the results obtained in 1998.

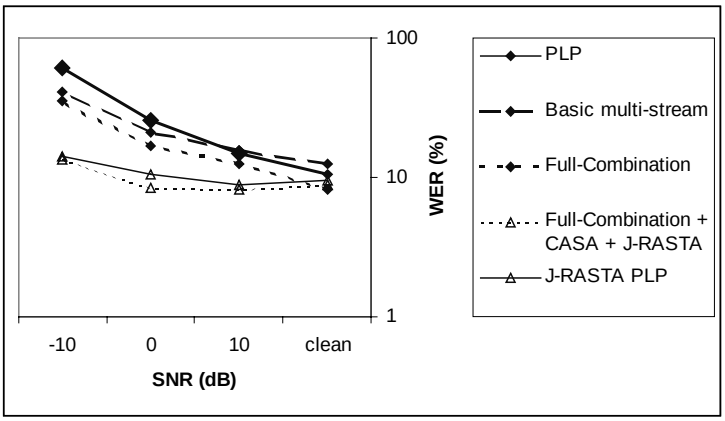


Abbildung 14: *Some results obtained in 1998 for robust speech recognition.*

♢ ARTIST – Articulatory Representations To Improve Speech Technologies

Funding: Swiss National Science Foundation

Duration: April 97 – March 99

Persons involved: Sacha Krstulovic and Georg Thimm

Description: This project involves both speech and vision groups. For the description of the project please refer to 3.2.5.

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

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

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

3.1.4 Speaker Recognition

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

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

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

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

- **Improving the client/world modeling**. In order to verify the identity claimed by a speaker, two stochastic models are generally used: one for the claimed customer and the other for the "world" (where all the speakers different from the client). Given an utterance from a user, the likelihoods of both the client and the world models are computed and the ratio is compared to a predefined threshold. In order to improve the robustness of such a system, the client and the world models must be improved. In 1998, several algorithms were developed in that direction at IDIAP and were shown to improve the speaker verification performance:

  1. *Synchronous Speaker / World Alignment*
     Both client and non-client (world) hypotheses are modeled with HMMs. The decision is then based on the maximum likelihood between the two HMMs given the observed utterance. Separation of the HMMs assumes that the two models are independent, which

is obviously not correct. A new structure is now being investigated where the speaker and the world are sharing the same HMM. We have established the theoretical foundation of such a model for optimal decoding and training. Figure 15 illustrates the principle of this new structure in comparison with the classical modeling.

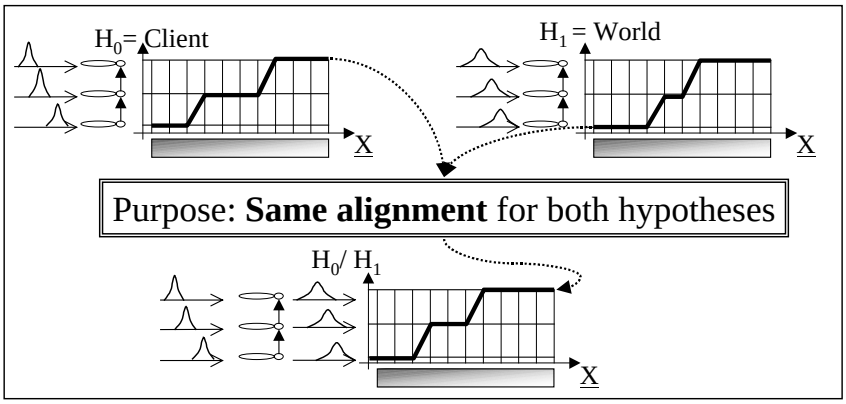


Abbildung 15: *Synchronous alignment approach.*

2. *Incremental Enrollment*
   HMMs adaptation algorithms can be used in an incremental enrollment strategy. As for speech recognition, adaptation allows to incrementally adapt the parameters of the models to better describe the data characteristics in the conditions of use. In the case of speaker recognition, these algorithms also allow to enrich the stochastic models initially trained on few enrollment utterances. These adaptation algorithms will also keep track of the inherent fluctuations of the customers' voices. Existing algorithms were studied and an API was defined for the implementation of these algorithms in the case of speaker verification.
3. *Customized Password*
   Text dependent speaker verification generally makes use of a password predefined by the system or the user. However, an added value can be obtained if the clients have the possibility to change their passwords easily. This requires the system to be able to infer the model of customized password. Several algorithms can be investigated in this context. These algorithms are similar to the ones developed for improving lexical modeling.

- **Improving the decision strategy**. Several issues may be studied to increase the robustness of the decision module. One important issue is the automatic selection of relevant acoustic vectors that yield robust discrimination between the customer and the world. Along this line, one approach has been investigated in 1998 regarding deliberate imposture. It consisted in automatically deriving imposture utterances that largely resemble the client utterances. Deriving such utterances can help adjusting the decision threshold and measuring the limitation of the state of the art speaker verification systems. A simple approach based on the concatenation of the client speech segments was investigated, and experimental results showed that current systems poorly resist such an imposture strategy.

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

the ELISA consortium, IDIAP participated in the 1998 NIST evaluation, and the fusion scheme resulted in leading edge performance.

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

♢ COST250 – Automatic Speaker Recognition over the Telephone Network

Funding: European project, COST action, supported by OFES

Duration: October 95 – September 98

Persons involved: Dominique Genoud, Johnny Mariethoz

Description: Several European laboratories from almost all the European countries participate in the COST action. The collaborative COST250 action aims at:

- studying the technological, economical and social feasibility of the use of automatic speaker recognition and speaker verification technologies in real life applications.
- analyzing in detail the applications in telecommunications.
- collecting the databases required for the development and evaluation of the algorithms.
- transferring the know-how between European laboratories.
- developing demonstration prototypes for these technologies.

Achievements: In 1998, in the framework of this COST250 project, IDIAP participated in the latest international NIST evaluation (National Institute of Standards and Technology, USA) and showed that their system was at the leading edge of speaker recognition technology. In collaboration with the ELISA consortium, IDIAP implemented a fusion algorithm of different systems and showed that this strategy was actually yielding significant improvement of the speaker verification performance.

♢ Enhanced automatic speaker recognition in telephony

Funding: Swiss National Science Foundation

Duration: April 96 – December 98

Person involved: Dominique Genoud

Description: This research project aims at performing more fundamental research in speaker recognition and speaker verification, including:

- analysis of intra- and inter-speaker variability, and selection of better parameters for speaker characterization.
- development of specific algorithms, more suitable to speaker verification tasks.
- development of adaptive environment techniques (noise, transmission channel, ...)
- development of decision taking from complementary cooperative tokens.
- evaluation of speaker recognition technology with regards to or in synergy with other biometric technologies.

Achievements: In 1998, several speaker transformation techniques have been proposed and developed. IDIAP has shown that this approach permits to deteriorate drastically the performance of a state of the art speaker verification system. In the framework of this project, M. D. GENOUD has terminated his PhD thesis with success.

♢ PICASSO – PIoneering Caller Authentication for Secure Service Operation

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

Duration: March 1998 – February 2001

Persons involved: Johnny Mariethoz, Dominique Genoud

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

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

Achievements: within the PICASSO project, IDIAP is mainly involved in the improvement of the state of the art speaker verification algorithms. In this context, IDIAP distributed the Polyvar database to the partners in 1998, participated actively in the setup of the reference system, and participated actively to the definition of the corresponding experimental protocol. IDIAP developed both the theoretical and software parts of the synchronous alignment algorithm described earlier. The results are shown in Figure 16. IDIAP defined the software API and algorithmic issues related to incremental enrollment.

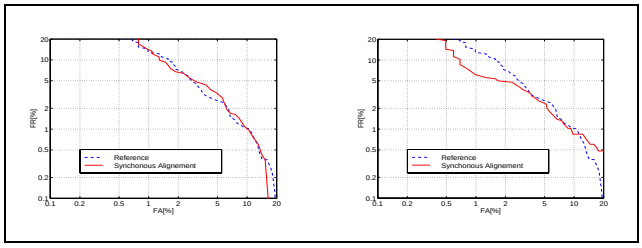


Abbildung 16: *Some results obtained in 1998 with synchronous alignment.*

3.1.5 Large Vocabulary Robust Speech Recognition

As compared to small/medium vocabulary speech recognition, Large Vocabulary Speech Recognition (LVCSR) requires an improvement of robustness at both lexical and linguistic levels, plus improvements in the interaction of all the layers. The models are more complex and have to be manipulated with care. In 1998, important work has been done at IDIAP to adapt the English hybrid HMM/ANN

systems to perform robust LVCSR in French. The following directions can be distinguished.

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

  In 1998, we have been focusing on hybrid systems. A large number of experiments have been conducted in order to determine the optimal size of the MLPs. It has been noticed that significant improvements could be obtained from **enlarging the size of the MLP**. Furthermore, an alternative way of using the **training** of the MLP has been developed and provided satisfactory results.

- Introducing **pronunciation variants**. Multiple pronunciations for some of the vocabulary enriches the lexical modeling. The variants can be computed automatically for each word separately or using some rules and inference techniques. The rules are generally described using a tree or a network, which can be stochastic. Stochastic trees offer an elegant description of the rules augmented with probability values. Several approaches have been studied in 1998 and have shown that improving lexical modeling highly increases the global performance.

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

- **Interaction between the modeling layers**. Interaction between the acoustic, lexical and linguistic levels forms a main research topic. Generally, the computed a posteriori score is a function of the acoustic likelihood, the lexical transcription probability and the language probability. However, it is appropriate to weight differently these contributions given (1) the hypotheses made during the estimation of the local likelihoods, and (2) the lack of balance and the difference in reliability of the different models. To improve the layers' interactions, the use of an **acoustic/language-scaling factor** was investigated. This factor weights differently the acoustic and language components in the global score. A good choice of such factor leads to significant improvements, as shown by the experiments conducted in 1998.

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

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

As a summary of the large LVCSR improvements achieved in 1998, we can evoke the results obtained on the BREF database: starting from a WER of 33%, using the previously described approaches resulted in a 23% WER. This represents a **relative improvement of 30%**.

All the preceding approaches have been studied and developed within the framework of several projects. These projects are detailed in the following. Important developments were also developed within the THISL and SWISSCOM projects that are described in sections 3.1.6 and 3.1.7 respectively.

◊ COST249 – Automatic Speech Recognition over the Telephone

Funding: European project, COST action, supported by OFES

Duration: October 95 – September 98

Persons involved: Giulia Bernardis, Johan Anderson

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

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

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

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

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

Achievements: In 1998, IDIAP has largely improved its LVCSR performance. This was achieved by

- optimizing the MLP size and MLP training
- using multiple pronunciations to describe words in the lexicon
- preprocessing the text databases used for training the language models
- replacing the bigrams with trigrams as language models.

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

3.1.6 Voice Thematic Indexing

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

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

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

Information retrieval algorithms measure the distance between a document and a particular request. Several decision criteria can be used: the simplest one is to find the requested word in the document. More elaborate stochastic retrieving algorithms compute some semantic distance between a document and the requested words. IDIAP has been working on this last class of algorithms. In 1998, IDIAP applied the basic **Latent Semantic Analysis (LSA) algorithm**, for retrieving purpose. LSA permits to infer a semantic cardinal continuous space where both important words and documents are represented as points. IDIAP extended the classical LSA algorithm using Self Organizing Maps (SOM) classifiers. One important problem, relative to retrieving information, is how to evaluate the different algorithms. IDIAP has proposed a new quantitative measure to **evaluate the retrieving algorithms**.

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

♢ THISL : THematic Indexing of Spoken Language

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

Duration: February 97 – January 2000

Persons involved: Johan Anderson, Mikko Kurimo, Djamila Mahmoudi

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

The expected result of the project is a real-time prototype system for navigating in the sound-track of a TV news broadcast. Significant intermediate results will include:

- transcription of broadcast speech
- development of audio editing tools
- content-based retrieval from audio/video archives
- a robust spoken language interface for search and retrieval of multimedia data.

Achievements: In 1998, IDIAP has largely improved its LVCSR performance for French recognition as shown in the section 3.1.5. IDIAP has also adapted the THISL prototype to demonstrate the retrieving capacities on a French voice database. At the research level, IDIAP has developed several variants of the LSA algorithm for retrieving information and has shown the usefulness of the perplexity measure to evaluate and to compare the retrieving algorithms. IDIAP has studied several algorithms for the speech/music detection problem.

3.1.7 Prototyping and Spoken Language Resources

Prototyping and spoken language resources are necessary to support technology developments. Building prototypes permits to observe the technology in a real-life environment. By the same way, it allows to understand its relative added value and weak points to improve. Finally, it permits to measure the acceptability of the technology by end users. As a matter of fact, no technology development and evaluation can be done without spoken language resources. For these reasons, IDIAP has a main interest into these two components of the technology development.

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

Spoken language resource developments were performed within two projects: AVIS/SWISSCOM project and the SPEECHDAT European project. These projects are described in more detail in the following.

♢ AVIS (SWISSCOM) – Advanced Vocal Interfaces Services

Funding: Swisscom

Duration: 1 year, January – December 98

Persons involved: Johan Anderson, Olivier Bornet, Andrew Morris

Description: The main goal of this project was to provide Swisscom with the necessary baseline technology to develop advanced vocal interface services. This project involved several research and development directions, including:

- Robust speech recognition based on classical HMMs and hybrid HMM/ANN systems.
- Development of demonstrator prototypes (voice dialing, personal attendant).
- Support of the ISIS project, especially for the interaction between acoustic decoder and linguistic models. This work is done in close collaboration with EPFL.
- Management and distribution of the Swiss-Polyphone database, with extension to and preliminary testing on GSM data.

Achievements: The results concerning robust speech recognition and interaction between acoustic and linguistic layers are described in the sections 3.1.3 and 3.1.5. Concerning prototyping, Voice dialing and personal attendant applications have been developed.

At the functional level, the voice dialing system (or voice directory dialer) is a typical computer telephony tool that allows a user to place calls simply by speaking the name of (or a keyword associated with) a person in the directory. Voice dialing is a personal telephone service, and customers can easily customize their system by defining their own directory. At the technological level, the approach considered at IDIAP is based on hybrid DTW/ANN systems. Limited keyword spotting capabilities have been added to the system, improving its robustness to noise. At the technical level, an XTL/SPARC Interactive Voice Response server was used for the CTI interface. A software library was developed to dynamically manage the data (references and directory) associated with each enrolled customer, and to allow enrollment of new customers.

Voice dialing is useful for outbound calls. But in parallel, inbound calls can be managed using a personal attendant, based on speaker independent speech recognition and a lexicon represented in terms of phonetic transcriptions (and possibly complemented with an email address associated with each entry). At the functional level, the persons calling the personal attendant are invited to pronounce the name of the person/department they want to get connected to. After recognition of the pronounced name, the server connects the external call to the internal number. At the technological level, the recognition core used here is based on the hybrid HMM/ANN technology. At the technical level, no specific architecture is required to maintain the users' data. Only the names of the persons attached to the personal attendant as well as their email addresses are necessary.

SWISSCOM databases have been extensively used. Labeling or formatting errors were detected and documented. In this framework, IDIAP developed a tool for the collection of the errors encountered by the users of these speech databases.

♢ SpeechDat – Spoken Language Resources Dissemination

Funding: European project LE2-4001, Telematics Program, supported by OFES

Duration: March 96 – February 98

Person involved: Frank Formaz

Description: Due to the progress reached nowadays in speech processing technology, more and more powerful voice driven teleservices can be implemented. They allow, for instance, easy access to information services (e.g. train table information), transaction services (e.g. home shopping), and call processing services (e.g. voice mail handling) via the telephone network. Many European companies are active in the field of creating such services and delivering the required speech technology. However, for research purposes and for the implementation of the speech processing technology (speech recognition and speaker verification), spoken language resources are necessary, including speech databases, lexica, and related tools.

The current project aims at producing, standardizing, evaluating and disseminating large speech databases covering a wide range of languages (most of the European languages) and applications. Due to the importance of language resources, numerous partners are involved in the project: Aalborg University (DK), British Telecom (UK), European Commission (L), CSELT (I), Tampere Univ. of Technology (FIN), ELRA (F), GEC-Marconi Ltd (UK), GPT Ltd (UK), IDIAP, INESC (P), Knowledge S.A. (GR), Kungl Tekniska Hogskolan (S), Lernout & Hauspie Speech Products (B), Matra Communication (F), Philips (NL), Philips

(D), Portugal Telecom (P), Siemens AG (D), Speech Processing Expertise Centre (NL), Swisscom (CH), Telenor R&D (N), Univ. of Maribor (SL), Univ. München (D), Univ. of Patras (GR), Univ. Politecnica de Catalunya (E), Vocalis Ltd (UK). In the framework of this project, IDIAP participates in the specification of the databases and is responsible for collecting and labeling the data in Swiss French and Swiss German in collaboration with ETH. This project has been terminated this year.

Achievements: In 1998, IDIAP released the SpeechDat database, which represents a subset of the Swiss-French Polyphone database satisfying the European Speechdat format. Its distribution has been entrusted to ELRA.

3.1.8 Software Development

As exposed above, IDIAP has participated to the development of several general-purpose software tools in collaboration with other laboratories, mainly in the framework of European projects. Towards the end of 1998, an important development activity has been initiated at IDIAP in order to write completely new IDIAP software, with the goal of supporting our general and specific research and development needs. A team involving researchers from different groups has been set up to develop and cross-validate the new software. Several modules have already been written. This development activity aims at creating a system that integrates both classical and hybrid HMM systems. It will handle small and large vocabulary recognition of isolated words and/or continuous speech. The new system will facilitate the integration of the new approaches developed at IDIAP.

3.1.9 Education

Education is also one of the missions of IDIAP. In 1998, IDIAP actively participated to several conferences and workshops. H. Bourlard, Director of IDIAP, teaches different courses at the Swiss Federal Institute of Technology at Lausanne (EPFL). Besides these activities, IDIAP actively participates (as the only official Swiss partner) to the set up of the European Masters in Language and Speech in the framework of a SOCRATES/OFES project.

♢ SOCRATES – European Masters in Language and Speech

Funding: European Project, DG XXII

Duration: September 97 – September 2000

Person involved: Herve Bourlard

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

3.2 Computer Vision Group

Group Leader: Juergen Luettin

Our objectives are to address research topics in the automatic analysis and interpretation of visual scenes to develop technologies and prototype applications that respond to the future need of the information society. Through activities in various research projects, the group has acquired expertise in the following major research areas:

- Object detection and recognition
- Motion analysis and recognition
- Shape analysis and representation
- Sensor fusion
- Document analysis and recognition

These research activities were often motivated or driven by particular applications in the areas of

- Multimodal interfaces
- Access security
- Document processing

which fall into the genenal application themes defined by IDIAP. Several projects have been investigated in collaboration with the speech processing and machine learning groups, taking advantage of different disciplines and complementary expertises with the aim of completely covering a scientific domain. Some projects are conducted in collaboration with industries or public organisations to validate the technology and to promote its exploitation.

3.2.1 Object Recognition

Object recognition represents one of the fundamental problems in computer vision and addresses the recognition and classification of visual objects. A sub-problem is concerned with the detection of objects of known classes. We have investigated and developed an algorithm for the fast detection of objects in unconstrained images, which is based on an artificial neural network (in our case, a multilayer perceptron) and which exploits the computational efficiency of an FFT. In comparison to conventional approaches, the method leads to a speed-up factor of about 8 - 16, depending on image and object size, while retaining identical performance levels.

Face Detection

Object detection techniques have been applied to the problem of face detection, which is a pertinent, yet difficult, task that is required in several vision applications, e.g. face recognition, visual speech recognition, image and video indexing, video conferencing, and video coding. The task of face detection consists in the analysis of an entire image aiming to detect all faces that appear in the image. It implies that the system does not know how many faces are present. Our face detection algorithm has been tested on difficult test images of natural scenes, containing several subjects, and has shown good performance of the system. The algorithm has been implemented as a demonstration system and is able to detect faces in real-time. A typical example is illustrated in Fig. 17.

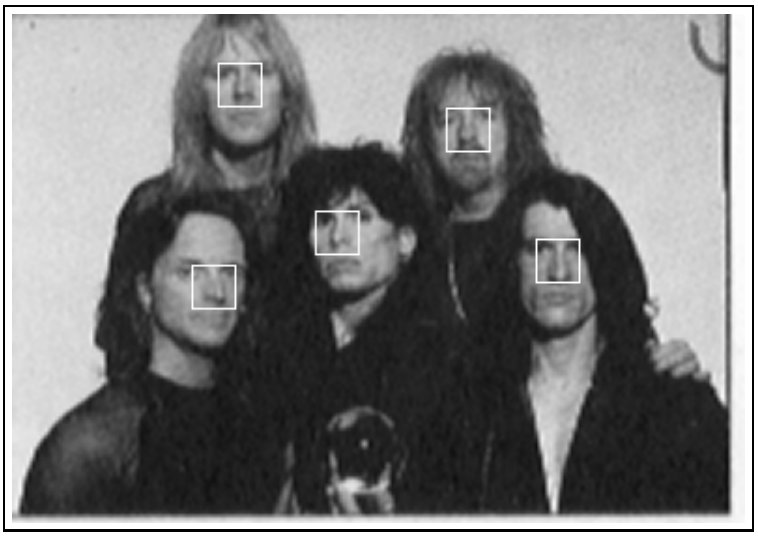


Abbildung 17: *Example of face detection results on an image of the band Aerosmith.*

Image Indexing and Retrieval

We are collaborating with the CVFP (Centre Valaisan du Film et de la Photographie) in the area of image indexing and retrieval. The objective is the automatic content-based indexing and retrieval of an archive of ancient black and white photographs that have been digitised by CVFP and that will be made publicly available on the WWW through the Bibliothèque Cantonale du Valais. The content based analysis of these images is very difficult since colour information, one of the most important image features used in current image retrieval approaches, is not present. A higher level image feature and search criterion of particular interest is the presence of persons in the photographs. Our first efforts have been investigating the content analysis by detecting faces in the photographs and the use of this information for image indexing and retrieval.

3.2.2 Audio-Visual Person Verification

The combination of several modalities for personal identity verification is motivated by the fact that monomodal systems often don't meet the high performance requirements imposed by typical applications, whereas the combination of modalities can improve their performance. Within the European project M2VTS we have made several contributions in the area of multimodal person verification. This project targets authentication applications in tele-services and access security and is of considerable commercial interest. It involves the companies Matra-Nortel (France), Cerberus AG (Switzerland), and Ibermatica SA (Spain)), and the potential end-users Compagnie Européenne de Télésécurité (France), Banco Bilbao Vizcaya (Spain), and Unidad Tecnica Auxiliar de la Policia (Spain). A typical setting for multimodal access control is depicted in Fig. 18.

Verification Modalities

We have developed a new method for person authentication that simultaneously combines the acoustic speech signal with visual motion information of the face while the person is talking. Visual information is extracted by tracking the lips over image sequences and extracting both shape and grey-level information of the mouth region. The verification technique uses a combination of acoustic-temporal and spatio-temporal modes based on hidden Markov models (HMM) (Fig. 19). As discussed in Sec. 3.1, HMMs are extensively used in speech processing and represent a very efficient method to model the

Abbildung 18: *Multimodal person verification based on face, speech, and lip movements.*

statistical variation in both the temporal and the feature domain. We exploit these properties to model the visual appearance as well as the voice characeristics of the speaker. For the combination of the two modalities we use the Multi-Stream approach that can account for asynchrony which might occur between the two signals. In comparison to monomodal verification methods, this approach is more robust to impostor accesses.

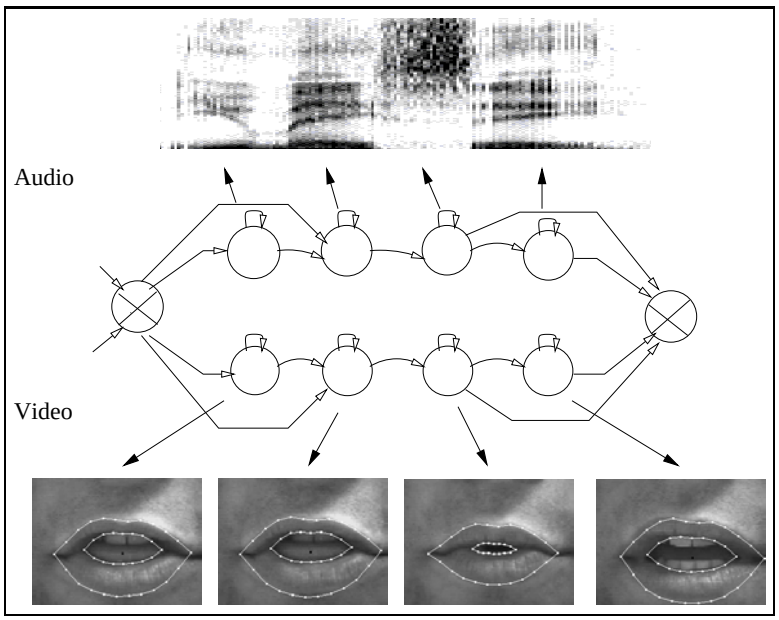


Abbildung 19: *Audio-visual speaker verification based on speech and lip movements.*

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

Sensor Fusion

Sensor fusion is a powerful solution to pattern recognition problems involving complementary classifiers and noisy input since it allows the simultaneous use of different information sources. Typical problems which arise in person verification research are the combination of different information sources (frontal face, face profile, facial motion, speech) and different information representations (still, dynamic). The fusion of classifiers is particularly difficult in the case of classifiers exhibiting different performance levels. The group has investigated and evaluated various methods of sensor fusion in collaboration with the Machine Learning Group, including Support Vector Machines (SVM), multi-linear classifiers, MLP, and C4.5. Experimental results have shown that the fusion of classifiers increases the performance and outperforms each single verification modality.

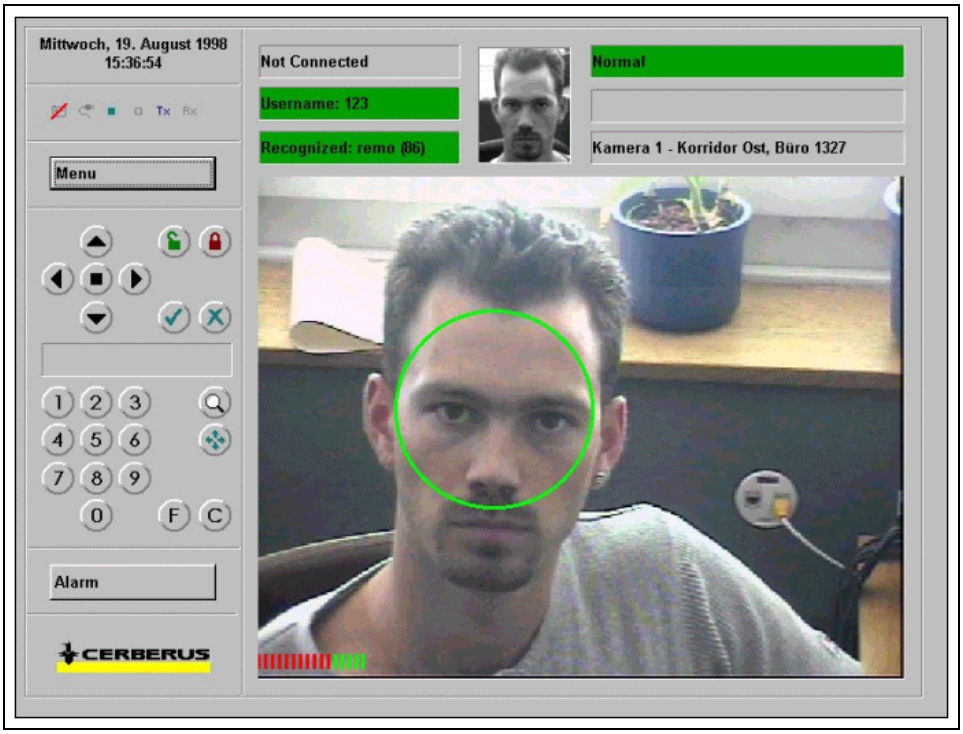


Abbildung 20: *Multimodal person verification for access control to secure buildings or restricted areas, implemented by Cerberus AG (Switzerland).*

Algorithm Evaluation

A large multimodal database of synchronised images and speech data of 295 persons has been collected within the M2VTS project and probably represents the largest database of its kind. The database has been used for the evaluation of algorithms by the different project partners. Our group has developed and evaluated two speech based verification modalities (text-dependent and text-independent) and one technique for classifier combination based on SVM. These techniques have obtained outstanding results in both monomodal and multimodal verification. In addition, we have tested our face detection algorithm on this database and have provided a quantitative performance measure of the method.

Demonstrator and Field Tests

The performance of verification systems tested on laboratory databases often drops considerably when

Abbildung 21: *Multimodal verification for access control to cash dispensers, developed for Banco Bilbao Vizcaya (Spain).*

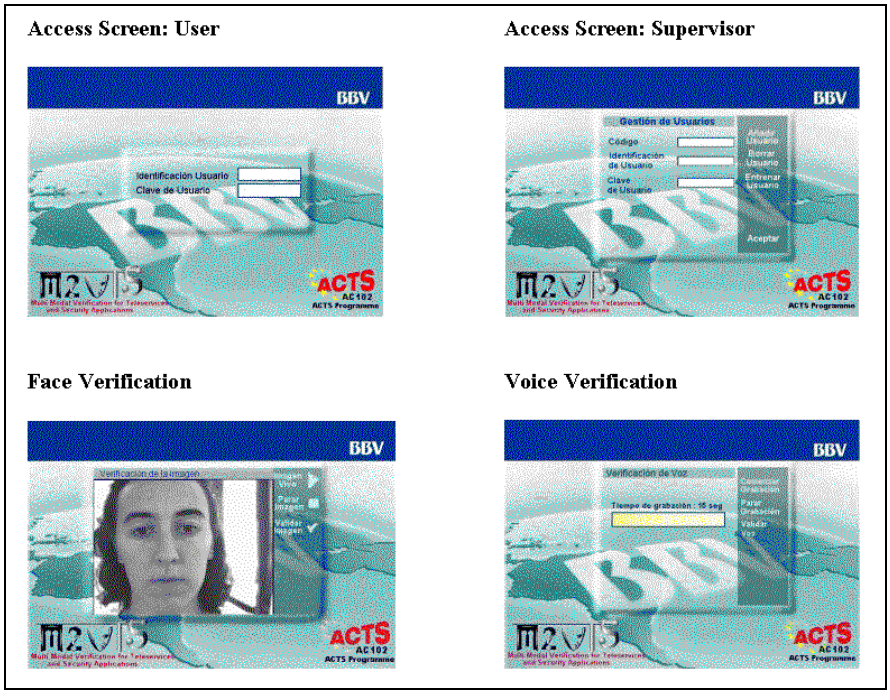


Abbildung 22: *Multimodal user verification for access control to restricted documents or services on the WWW (tele-banking, tele-shopping, tele-working, etc.).*

applied to real-world situations. In collaboration with Cerberus AG and the University of Neuchatel, we have implemented a verification system prototype including user interface, client registration, training, and verification possibilities. To allow the evaluation of verification performance in realistic applications, the system has been used to collect a field test database, where 22 members of IDIAP were recorded during a period of several months. This database represents a very difficult application scenario which is characterised by a low-cost microphone, many non-native speakers, reverberations, and varying distance and direction of the microphone. Verification experiments have been performed on this database to test the robustness of algorithms under these conditions. Potential applications targeted by Cerberus are access control to secure buildings (Fig. 20) and remote verification of triggered alarms.

Ibermatica SA, has integrated our verification technology and has developed several application prototypes including multimodal verification for cash dispenser access (Fig. 21), internet access (Fig. 22), and computer access.

♢ M2VTS – Multimodal Verification for Teleservices and Security Applications

Funding: European project AC 102, ACTS Program, supported by OFES

Duration: October 95 – November 98

Person involved: Souheil Ben-Yacoub and Gilbert Maître

Partners: Matra Communication (F), Cerberus AG (CH), Ibermática S.A. (E), Ecole Polytechnique Fédérale de Lausanne (CH), Université de Neuchâtel (CH), Université Catholique de Louvain (B), University of Surrey (GB), Renaissance (B), Aristotle University of Thessaloniki (GR), Compagnie Européenne de Télésécurité (F), Universidad Carlos III (E), Banco Bilbao Vizcaya (E), Unidad Tecnica Auxiliar de la Policia (E)

Description: The primary goal of the M2VTS project is to address the issue of secured access to local and centralized services in a multi-media environment. The main objective is to extend the scope of application of network-based services by adding novel and intelligent functionalities, enabled by automatic verification systems combining multimodal strategies (secured access based on speech, image and other information). The objectives are also to show that limitations of individual technologies (speech recognition, speaker verification...) can be overcome by relying on multimodal decisions (combination or fusion of these technologies) and can find practical and important applications in the new emerging fields of advanced interfaces for tele-services.

Achievements: We have developed a text-dependent and a text-independent speaker verification method. Both techniques have been evaluated on the M2VTS, XM2VTS, and LoCoMic databases. A new verification technique based on visual speaker modelling was developed and combined with an acoustic speaker verification system. Several classifier combination methods were evaluated at IDIAP including SVM, multi-linear classifier, and LAD (logical analysis of data). The monomodal and multimodal systems developed at IDIAP obtained the highest results on the XM2VTS database within the consortium. A face detection system was also developed to allow the unconstrained use of face recognition systems. We have implemented a PC-based speaker verification demonstrator that was used to collect the LoCoMic database of IDIAP staff on which the developed algorithms were evaluated. The text-independent speaker verification method was adapted and integrated for application demonstrators at Cerberus AG and at Ibermatica SA.

3.2.3 Audio-Visual Speech Recognition

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

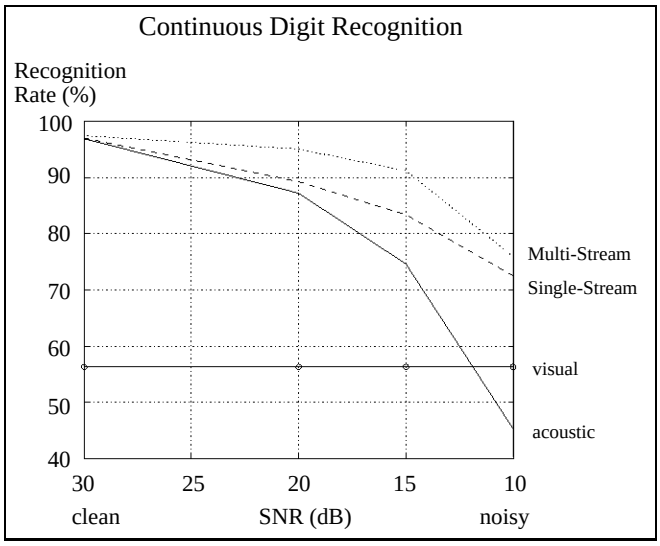


Abbildung 23: *Recognition rate for acoustic, visual, and audio-visual speech recognition.*

Audio-visual performance is displayed for traditional fusion at the feature level (Single-Stream) and for fusion based on the Multi-Stream method.

Appearance based modelling

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

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

Sensor Fusion

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

present, as illustrated in Fig. 23.

♢ AV-COM – Audio-Visual Combination

Funding: Swiss National Science Foundation, "R'EQUIP" program for new equipment.

Duration: October 97 – September 98

Description: New computer facilities (CPU server and real time audio-video acquisition with audio-video synchronization) for research into multimodal audio-video processing.

The objective of the present R'EQUIP research grant is to provide IDIAP with the necessary computer resources to carry out multimodal signal processing research. This new equipment will support the realization of new projects in multimodal data processing and accelerate the accomplishment of existing projects in the audio-visual domain. These projects represent individual modules in the area of multimodal human computer interfaces.

3.2.4 Facial Expression Recognition

The objective of this project is to develop robust and accurate computer vision techniques for the visual analysis and recognition of facial expressions from image sequences. Figure 24 displays example images showing different facial expressions. Facial expressions recognition should not be confused with human emotion analysis as it is often done in the computer vision community. Whereas facial expression recognition is concerned with the classification of facial motion into abstract classes, purely based on visual information, human emotion is the result of many different factors and it's state might (or might not) be revealed through a number of channels of which the face is just one of them.

Abbildung 24: *Example images showing four different facial expressions.*

The application of currently available automatic facial expression recognition systems to the analysis of natural scenes is often very restricted due to the limited robustness of these systems and the hard constraints posed on the subjects and on the recording conditions. Our approach aims to overcome these shortcomings by investigating advanced computer vision techniques for the modeling of faces and scenes. It includes appearance based models for object localization, parametric model based techniques for motion estimation, and stochastic finite state machine methods for the temporal modeling of facial actions.

This project benefits from collaboration with the Psychology Department at the University of Geneva, which has considerable expertise in the area of facial action coding and human emotion analysis. We are also collaborating with the Department of Otolaryngology at the Geneva University Hospital that is interested in the automatic evaluation of facial nerves.

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

Funding: Swiss National Science Foundation

Duration: October 98 – September 00

Partners: University of Geneva

Person involved: Souheil Ben-Yacoub and Beat Fasel

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

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

Achievements: The first few months of the FaxeX project have mainly been spend to conduct an extensive literature survey about the state-of-the-art in facial expression recognition and emotion analysis. Together with the Psychology Department at the University of Geneva we have defined a possible methology for the quantitative evaluation of automatic facial expression recognition systems and with the University Hospital of Geneva we have targeted a possible application.

3.2.5 X-Ray Image Sequence Analysis

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

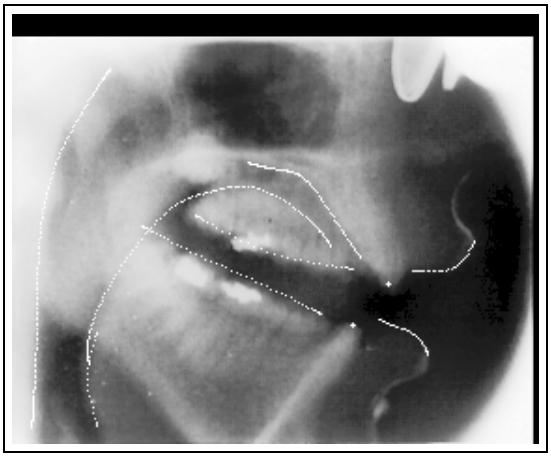


Abbildung 25: *Extraction of articulatory parameters form X-ray image sequences.*

A technique has been developed within the group to track tongue, lips, teeth, and throat in the X-ray film database provided by ATR (Japan), which is probably the largest database of its kind. Our tracking method uses specialised histogram normalisation and a tracking method that is robust against occlusion, noise, and spontaneous, non-linear deformations of the articulators. The tracking robustness is improved by the introduction of temporal constraints, which, however, can cause tracking errors in the case of fast movements. To compensate for this effect, a method that tracks the articulators both

forward and backward in time and that optimally combines the results, has been developed. Figure 25 shows an example X-ray image together with the extracted articulators.

♢ ARTIST – Articulatory Representations To Improve Speech Technologies

Funding: Swiss National Science Foundation

Duration: April 97 – March 99

Person involved: Sacha Krstulović and Georg Thimm

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

Subtasks involved in this research include :

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

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

3.2.6 Document Analysis and Recognition

Handwritten Character Recognition

Document analysis and recognition is concerned with the recognition of machine-printed, hand-printed, and hand-written documents. Research activities in our group started in 1992 and since then mainly concentrated on handwritten text recognition. The effectiveness of our technology has been demonstrated at the international NIST (National Institut of Standards and Technology) evaluation on U.S. census forms in 1994, where our system obtained the second highest performance level. The recognition is based on extracting a large number of character sub-images from the input that are individually classified using a MLP. The viterbi algorithm is used to assemble the individual classified character subimages into optimal interpretations of the input. These results are integrated with the use of a dictionary and a language model.

Recent work has concentrated on hand-printed character recognition, namely visual feature representations and similarity measures. Different feature representation methods have been investigated and were compared to traditional approaches. Work in the area of classification has been investigating the use of Support Vector Machines (SVM) and of MLPs for the classification on handwritten digits (Fig.26). SVM have recently gained much interest and have shown several advantages over alternative classifiers. One of the drawbacks of the SVM technique though is that it is a binary classifier that can not easily be applied to multi-class problems. The Machine Learning group has proposed a technique that allows the application of SVM for multi-class classification such as character recognition.

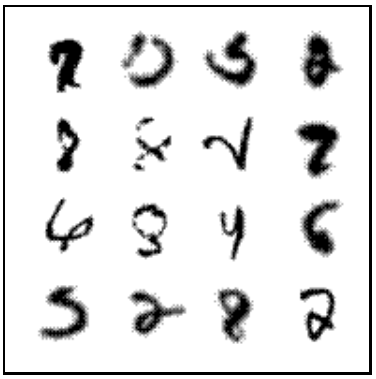


Abbildung 26: *Examples of handwritten digits, taken from the NIST Special Database 3.*

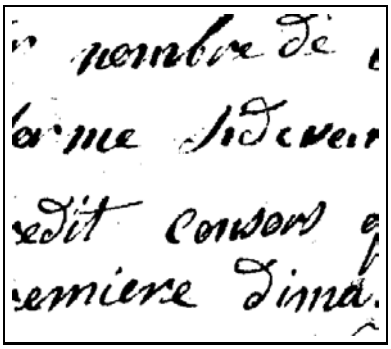


Abbildung 27: *Example of ancient handwritten text, taken from the Federal Archives of Switzerland.*

Cursive Handwriting Recognition

One of the fundamental problems in cursive handwriting recognition is the segmentation of characters. This problem is similar to the segmentation of phones in speech recognition and can be addressed by joint segmentation and recognition using stochastic methods. This strategy also allows the incorporation of multiple knowledge sources such as lexica, language models, and topic of the document. As application, we are targeting the recognition of ancient script documents (Fig. 27), which represents a particularely difficult application of cursive handwriting recognition.

3.3 Machine Learning Group

Group Leader: Eddy Mayoraz

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

The activity of our group is always application driven. The learning techniques involved are not studied for their own sake, but are always targeted towards a specific application. As illustrated in Figure 9, the research themes in the Machine Learning Group are articulated along two lines :

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

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

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

Thus, our research effort as support for the Speech Processing Group and the Computer Vision Group involves being aware of the latest learning techniques (e.g. *mixtures of experts*, *support vector machines*) and working on strategies to adapt any efficient learning technique to very large problems. The decomposition of large problems into a series of simpler subproblems provides a general approach that does not depend on the learning technique to be used and which has been shown to be quite efficient in practice.

Logical analysis of data (LAD) is a recent method aiming at extracting knowledge from data in a form that is as easy to understand as possible. This theory has emerged in the operations research community and is still not well known in the learning community. Almost all the work related to LAD is so far carried out at RUTCOR (Rutgers Center for Operations Research) and at IDIAP, involving a fruitful collaboration between the two research groups.

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

3.3.1 Divide and learn

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

- each subproblem is simpler than the global problem;
- some redundancy in the decomposition makes the global model more robust;

- both in the training stage and in its usage, the process can easily be parallelized.

K-class classification turned into 2-class classification

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

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

Figure 28 illustrates this idea on a simple example of 10 classes decomposed into 4 dichotomies. Each dichotomy is represented by a closed line whose inside and outside define the two classes. This example

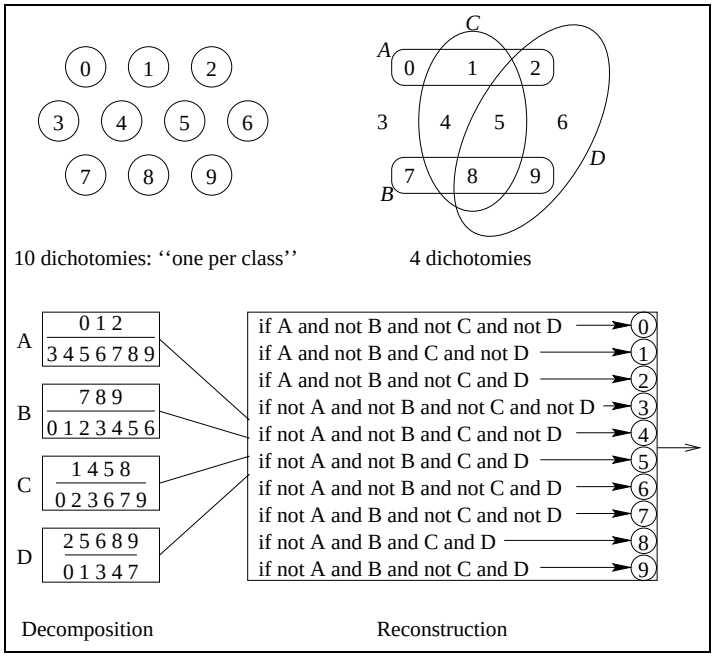


Abbildung 28: *Decomposition of a 10-class problem into 10 dichotomies (up left) and 4 dichotomies (up right).*

can be refined in many respects. The rule-based reconstruction pictured in the lower part of figure 28 can be advantageously replaced by an algebraic expression (vector-matrix product). An alternative would be to consider the reconstruction as a new classification problem for which any method could be investigated (stacking). The number of subtasks can be enlarged on purpose in order to get additional robustness from the reconstruction. Moreover, in the example of figure 28, each class is involved in each subtask to determine either positive or negative examples. But generally, some classes can be ignored by some subtasks (e.g. a subtask is defined for each pair of classes, discriminating between these two and ignoring the others).

For two years now, our group has carried out important work in this field. Our main contributions are:

- strategies to elaborate the decomposition by taking into account the distribution of data in the input space (a posteriori decomposition);

- various reconstruction methods (a posteriori reconstruction);
- iterative elaboration of the decomposition merged with the reconstruction;
- handling the reconstruction when the methods used to learn the subtasks have non-normalized outputs, (e.g. *support vector machines*).

Mixture of experts

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

One model known as *mixture of experts* and proposed in 1991 by Jacobs, Jordan, Nowlan and Hinton, is exactly of that form. Often presented as a generalization of neural networks, it is composed of *experts* realizing the subtasks and a *gating* determining the coefficients of a linear recombination of the outputs of the experts, as illustrated in Figure 29. Research on mixture of experts models has

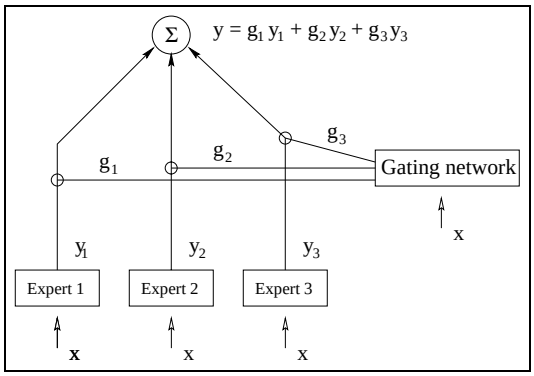


Abbildung 29: *A mixture of three experts.*

been carried out in our group within the following project.

♢ Hardware Friendly Neural Networks

Funding: Swiss National Science Foundation, FN 21-45621.95

Duration: April 96 – March 99

Partners: Swiss Federal Institute of Technology (EPFL)

Person involved: Perry Moerland

Description: The original aim of this project was the study and the optimization of artificial neural networks in order to ease their hardware implementation. According to the original plan, this improvement had to be done at three levels : the neuron (considering the neural function effectively realized by an analog implementation), the interconnection (understanding the influence of discrete weights with a low quantization level simulating a cheap numerical storage), and the topology (highly interconnected networks are not suitable for hardware representation).

Achievements: The first two aspects have been considered during the first year of the project and now the focus has been placed on the third one. More precisely, the research effort is concentrated on the study of a modular neural network model, known as *mixtures of experts*. This model adaptively partitions the input space (using a *gating* network) and attributes local experts to these regions.

Research within this project has focused on classification problems and the choice of the gating network using methods for density estimation, especially mixture models. It has been shown that mixtures of latent variable models are a more flexible alternative for Gaussian mixture models on the problem of input density estimation. However, including these mixture models as a gating network in a mixture of experts did not lead to better results than the ones with standard mixtures of experts (with a single or multi-layer perceptron gate). We are currently investigating how Bayesian techniques can be applied to a mixture of experts in order to control model complexity in a systematic way.

Expert fusion

When a classification problem is decomposed into several subproblems (see above), it has been demonstrated that it is often interesting to use different learning techniques for the resolution of the subproblems. This also holds for simple classification problems. Even in the case of a 2-class classification problem, several subproblems can be created. This can be based on different features (see 3.2.2), on a resampling of the training data (bagging, boosting, arcing), or using different types of models . The problem is then to fuse different decisions coming from each classifier into one final decision. In the most interesting case, the decision of each classifier is not just hard decision of class membership, but it also includes a confidence measure or a probability for this membership.

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

3.3.2 Learn and understand what you learn

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

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

However, while the human brain is comfortable with IF-THEN-ELSE rules, it is not common to cascade more than two or three of them. A more appealing approach, as far as human understanding is concerned, consists in enumerating a long list of indicators (or patterns or syndromes) that are not strict rules, but just arguments in favor or against a particular diagnosis. These indicators, are not embedded into a hierarchical structure, on the contrary, they are aggregated in a single weighted sum. Given a new case, if many patterns in favor of diagnosis A are active while only few patterns in favor of other conclusions are active, then the diagnosis A is chosen.

Logical Analysis of Data (LAD) is a new learning technique which is based on this idea. The elaboration, development and extension of this method is mostly done at RUTCOR and at IDIAP. Among other things, at IDIAP we are maintaining and evolving software for the experimentation of LAD. In 1998, a special effort was made in the development of a user-interface on a Windows

platform. Figure 30 pictures the part of this interface for the design of one batch execution of the method. Besides these software activities, we are working on several improvements of the current

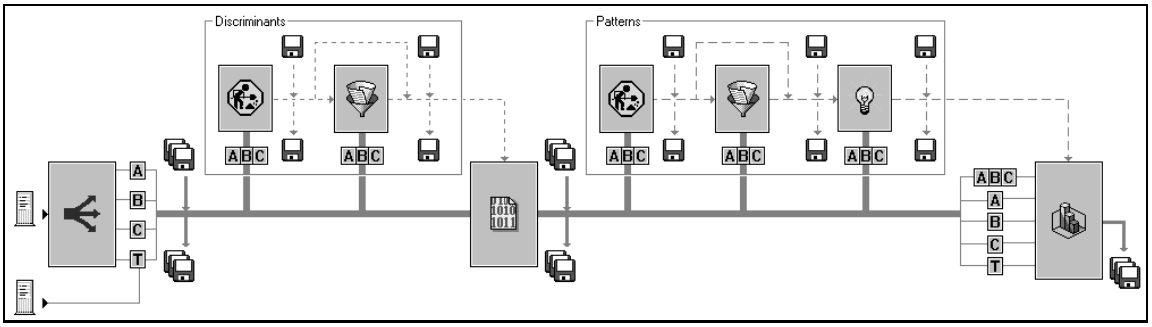


Abbildung 30: *View of the lab where batch executions of LAD are set up.*

The data flow is going from left to right. In the lower line, the first module is splitting the the data into different training sets (A, B, C) and a test set T; the second module is binarizing the data and the third module is measuring the performance of the model. The binarization is based on a set of discriminants, generated and then pruned by the first two upper modules. The model itself consists of a set of patterns, generated, pruned and weighted by the last three upper modules.

method in order to be able to deal with large scale problems. This research is supported by the following grant of the Swiss NSF.

♢ GLAD – Generalization of LAD

Funding: Swiss National Science Foundation, FN 21-46 974.96

Duration: November 96 – October 98

Partners: Swiss Federal Institute of Technology (EPFL)

Person involved: Miguel Moreira

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

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

Achievements: Within the first year of the project, an important work related to the decomposition of K-class problems into a series of dichotomies (see Section 3.3.1) was achieved in order to extend LAD to more general classification problems. So far, this problem was considered independently of the learning method used to solve each dichotomy. Some work is still to be done to specialize the results to the case of LAD.

Another achievement is related to the binarization process (first phase of LAD), where the complexity of the previous algorithm in $O(N^2)$ has been reduced to $O(N \log N)$, where N is the number of data items. This allows us to deal with much larger data sets.

3.3.3 Time series prediction and modeling

The set of problems addressed in machine learning has a coarse division into two classes: classification and regression, depending whether the output of the system takes its values into a finite unordered set or an ordered set. A first refinement of this taxonomy is obtained with the distinction between problems with static behavior (the output of the underlying system depends only on its input) and those with dynamic behavior (the output of the system at a given time depends on its current input as well as on the input history). In a first approximation, one could argue that the second type can be reduced to the first type by expending the input of the system to a window on the most recent inputs. However, this is not suitable for many applications, as it is too sensitive to time-scale variation, and to insertion and deletion of short events.

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

♢ ZEPHYR – Time Series Prediction with Hybrid Markov Models

Funding: Swiss National Science Foundation, FN 21-50744.97

Duration: January 98 – December 99

Partners: Swiss Federal Institute of Technology (EPFL)

Person involved: Frédéric Gobry

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

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

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

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

3.3.4 Spatial data analysis

A second prospecting activity of our group is devoted to the analysis of spatial data. This presents some analogies with TSP where the time is replaced by higher dimensional variables (2, 3 or even 4) but where interpolation usually replaces extrapolation.

The goal of our research is to exploit and adapt methodologies from the artificial neural network field for the analysis of environmental spatial data.

♢ CARTANN – Cartography by Artificial Neural Networks

Funding: Swiss National Science Foundation, FN 2100-054115.98

Duration: December 99 – December 2000

Partners: Lausanne University (prof. Michel Maignan)

Person involved: Mikhael Kanevski and Nicolas Gilardi

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

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

Final solutions proposed for the resolution of geostatistical problems will mostly be hybrids involving ANNs to extract the general trends, together with classical approaches of geostatistics such as kriging estimations and simulations to estimate the residuals of the ANN predictions.

Achievements: In the first month of work, some experiments were carried out using support vector machines (SVM) to discriminate between two classes of pollution in the Lake of Geneva (cadmium below or above a threshold). Figure 31 illustrates the effect of the parameter σ associated to RBF kernels of SVM.

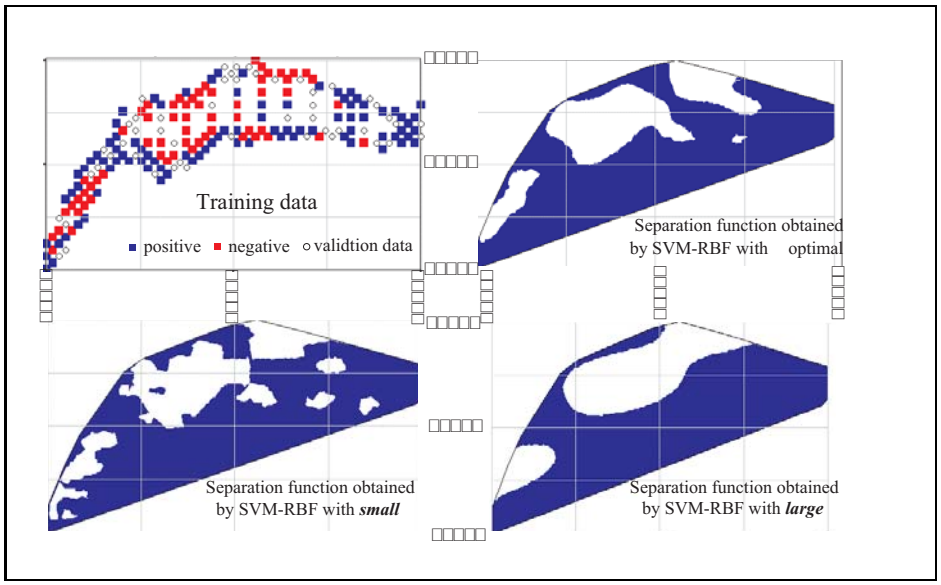


Abbildung 31: *Cadmium level in the Lake of Geneva.*

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

4 Educational Activities

4.1 Current Ph.D. Theses

- **Ph.D. Candidate:** Johan Myhre Anderson
  Supervisor: Prof. Hervé Bourlard
  Research topic: Robust Speech Recognition
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Giulia Bernardis
  Supervisor: Prof. Hervé Bourlard and Dr. Martin Rajman (EPFL)
  Research topic: Speech Recognition
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Beat Fasel
  Supervisor: Dr. Juergen Luettin and Dr. Souheil Ben-Yacoub
  Research topic: Facial Expression Recognition
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Nicolas Gilardi
  Supervisor: Prof. Michel Maignan (UNIL) and Dr. Eddy Mayoraz
  Research topic: Cartography using neural networks
  University: Lausanne University
- **Ph.D. Candidate:** Hervé Glotin
  Supervisors: Prof. Hervé Bourlard and Dr. F. Berthommier (ICP, Grenoble)
  Research topic: Coupling of CASA and Multistream recognition
  University: INPG, Grenoble
- **Ph.D. Candidate:** Frédéric Gobry
  Supervisor: Dr. Eddy Mayoraz and Prof. Hervé Bourlard
  Research topic: Time Series Prediction with Hybrid Markov Models
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Astrid Hagen
  Supervisor: Prof. Hervé Bourlard and Dr. Andrew Morris
  Research topic: Multistream Speech Recognition
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Katrin Keller
  Supervisor: Dr. Chafic Mokbel and Prof. Hervé Bourlard
  Research topic: Multichannel Speech Recognition
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Christopher Kermorvant
  Supervisor: Dr. Chafic Mokbel and Prof. Hervé Bourlard
  Research topic: Robust Speech Recognition

University: EPFL, Lausanne

- **Ph.D. Candidate:** Sacha Krstulović
  Supervisor: Dr. Chafic Mokbel and Prof. Martin Hasler (EPFL)
  Research topic: Using Articulatory Features for Speech Recognition / Speaker Verification
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Perry Moerland
  Supervisor: Dr. Eddy Mayoraz and Prof. Wulfram Gerstner (EPFL)
  Research topic: Mixtures of experts
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Miguel Moreira
  Supervisor: Dr. Eddy Mayoraz and Prof. Alain Hertz (EPFL)
  Research topic: GLAD – Generalization of LAD
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Bojan Nedić
  Supervisor: Prof. Hervé Bourlard and Dr. Chafic Mokbel
  Research topic: Speaker verification
  University: EPFL, Lausanne
- **Ph.D. Candidate:** Todd Stephenson
  Supervisor: Dr. Andrew Morris and Prof. Hervé Bourlard
  Research topic: Bayesian Networks applied to speech recognition
  University: EPFL, Lausanne

4.2 Ph.D. exams

- **Ph.D. candidate:** Dominique Genoud
  Supervisor: Prof. Martin Hasler (EPFL), Dr. Gérard Chollet (ENST-Paris)
  Examiners: Prof. Christian Wellekens (EURECOM), Dr. Régine Andre-Obrecht (IRIT-Paris).
  University: EPFL, Lausanne
  Title: Reconnaissance et Transformation de Locuteurs (Recognition and Transformation of Speakers)
  Short summary: The generic goal of the present PhD thesis is to understand how to analyse, decompose, model and transform the vocal identity of a speaker as seen through an automatic speaker recognition application, in view of improving current state-of-the-art speaker verification approaches. The Thesis starts with an introduction discussing the properties of the speech signal and the basis of state-of-the-art automatic speaker recognition systems. The errors of an operating speaker recognition application are then analysed. From the deficiencies and mistakes observed in a typical application, conclusions are drawn which imply a re-evaluation of the characteristic parameters of a speaker, and the modification of some parts of the automatic speaker recognition chain.

  Starting from the speech signal, the speaker characteristic parameters are extracted using an analysis and synthesis harmonic plus noise model (H+N). The analysis and re-synthesis of the harmonic and noise parts indicate those parameters which are speech or speaker

dependent. It is then shown that the speaker discriminant information can be found by subtracting the H+N modeled signal from the original signal.

A study of the impostor modeling, essential in the tuning of a speaker recognition system, is then carried out. The impostors are simulated in two ways. First by a transformation of the speech of a source speaker (the impostor) to the speech of a target speaker (the client) using the parameters extracted from the H+N model. This way of transforming the parameters is efficient as the false acceptance rate grows from 4% to 23%. Second, an automatic imposture by speech segment concatenation is carried out. In this case the false acceptance rate grows to 30%. A way to become less sensitive to the spectral modification impostures is to remove the harmonic part or even the noise part modeled by the H+N from the original signal. Using such a subtraction decreases the false acceptance rate to 8% even if transformed impostors are used.

To overcome the lack of training data (one of the main cause of modeling errors in speaker recognition), a decomposition of the recognition task into a set of binary classifiers is proposed. A classifier matrix is built and each of its elements has to discriminate between the data coming from the client and another speaker (referred to as "anti-speaker") randomly chosen. With such an approach, it is possible to weight the results according to the vocabulary or the neighbors of the client in the parameter (acoustic) space. The outputs of all the binary classifiers (matrix classifiers) are then combined according to a weighted sum to produce a single output score for each client input. The weights are estimated on an independent validation set to minimize the overlap between the client and impostors densities. It is shown that the binary pair speaker recognition system usually performs better that a state-of-the art HMM based system (especially in the case of a priori threshold).

In order to set a point of operation (i.e., a point on the COR curve) for the speaker recognition application, an *a priori* threshold has to be determined. Theoretically, the threshold should be speaker independent when stochastic models are used. However, practical experiments show that this is not the case and, due to modeling assumptions, the threshold actually becomes speaker and utterance length dependent. A theoretical framework showing how to adjust the threshold using the local likelihood ratio is then developed.

Finally, a further modeling error correction approach is proposed and tested using decision fusion. Practical experiments show the advantages and drawbacks of the fusion approach.

4.3 Student Projects

- **Trainee:** Alain Dannaoui

  School: EPFL, Lausanne

  Formation: semester project

  Subject: Analysis of Geo-Statistical Data with Neural Networks

  Duration: October 1997 – February 1998

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

- **Trainee:** Beat Fasel

  School: EPFL (EURECOM)

  Formation: diploma thesis

  Subject: Fast Multiscale Face Detection

  Duration: January 1998 - June 1998

  Supervisors: Dr. Souheil Ben-Yacoub, Dr. Jürgen Lüttin and Dr. Stephane Marchand-Maillet (EURECOM)

- **Trainee:** Ana Merchan
  School: University of Extremadura
  Formation: diploma thesis
  Subject: A posteriori reconstruction of decomposed classification methods
  Duration: October 1997 – March 1998
  Responsible: Dr. Eddy Mayoraz

- **Trainee:** Samuel Vannay
  School: EURECOM/EPFL
  Formation: project
  Subject: Réalisation d'un majordome vocal
  Duration: July 97 – January 98
  Responsible: Olivier Bornet and Prof. Giovanni Coray (EPFL)

- **Trainee:** Raphael Dupont
  School: EIV
  Formation: project followed by diploma thesis
  Subject: Développement d'une interface pour l'analyse logique de données
  Duration: April 98 – June 98 and October 98 – janvier 99
  Responsible: Dr. Eddy Mayoraz and Prof. Marylène Micheloud (EIV)

- **Trainee:** Fabrice Moret
  School: EIV
  Formation: project
  Subject: Construction de projections avec préservation des distances
  Duration: April 98 – June 98
  Responsible: Dr. Eddy Mayoraz and Prof. Gianni Pante (EIV)

- **Trainee:** Thierry Collado
  School: EIV
  Formation: diploma thesis
  Subject: Pilotage d'équipements par commande vocale
  Duration: octobre 98 – janvier 99
  Responsible: Olivier Bornet and Prof. François Corthay (EIV)

4.4 Lectures

- **Lecturer:** Prof. H. Bourlard
  Title: Decision, estimation, and statistical pattern recognition – Application to Speech Recognition
  Location: Pre-doctoral school for Computer Science and for Communication Systems, EPFL, Lausanne
  Duration: one semester from March 12 to June 18, 1998

4.5 Seminars

- **Title:** Utilisation de classificateurs binaires pour la classification multi-classe
  Speaker: Dr. Eddy Mayoraz
  Location: CUI, Geneva University
  Date: April 28, 1998
- **Title:** Decomposition of K-class classification problems into dichotomies
  Speaker: Dr. Eddy Mayoraz
  Location: INFM & DISI, Computer Science Department, Genova University
  Date: May 6 – 7, 1998
- **Title:** Mixtures of latent variable models for localized mixtures experts
  Speaker: Perry Moerland
  Location: MANTRA seminar, EPFL, Lausanne
  Date: May 14, 1998
- **Title:** Utilisation de classificateurs binaires pour la classification multi-classe
  Speaker: Dr. Eddy Mayoraz
  Location: MANTRA, EPFL, Lausanne
  Date: May 22, 1998
- **Title:** Some solutions to the missing feature problem in data classification, with application to noise robust ASR
  Speaker: Dr. Andrew Morris
  Location: Signal processing laboratory (LTS), EPFL, Lausanne
  Date: June 4, 1998
- **Title:** Phoneme transitions in ASR - Why Use Them?
  Speaker: Dr. Andrew Morris
  Location: Signal processing laboratory (LTS), EPFL, Lausanne
  Date: June 4, 1998
- **Title:** Traitement de la parole et communication parlée
  Speaker: Prof. H. Bourlard
  Location: Ecole d'été Européenne – Communication Homme-Machine, Alpe d'Huez
  Date: September 10, 1998
- **Title:** Automatic speech and speaker recognition by machines: a technical and practical overview
  Speaker: Prof. H. Bourlard
  Location: Computer Science Postgraduate Course on Multimodal Interfaces, EPFL
  Date: September 18, 1998
- **Title:** Génération de support vecteurs: une méthode mathématique pour l'analyse de données
  Speaker: Dr. Eddy Mayoraz
  Location: DMA and DI, EPFL (Colloque en mémoire du professeur Charles Rapin)

Date: October 6 1998

- **Title:** Speaker recognition at IDIAP

  Speaker: Dominique Genoud

  Location: University of Surrey, UK

  Date: October 13 1998

- **Title:** Speaker recognition: from laboratories to applications

  Speaker: Dominique Genoud

  Location: British Machine Vision Association (BMVA), British institute of radiology and Roentgen Institute, London, UK

  Date: October 14 1998

- **Title:** Research at IDIAP and Multi-Stream Processing

  Speaker: Prof. H. Bourlard

  Location: Computer Science, ETH, Zürich

  Date: November 20, 1998

- **Title:** Hybrid systems based on hidden Markov models for time series analysis: preliminary study for the case of exchange rates.

  Speaker: Frédéric Gobry

  Location: Olsen Associates, Zürich

  Date: November 20, 1998

4.6 Examinations

- **School:** EPFL – LSL

  Subject: Diploma work

  Expert: Mr. Perry Moerland and Dr. Gilbert Maître

  Candidate: Christophe Bregand

  Title: Conception d'un simulateur de réseau de neurones

  Date: March 11, 1998

- **School:** EPFL

  Subject: PhD thesis committee

  Expert: Prof. Hervé Bourlard

  Candidate: Jean Hennebert

  Title: Hidden Markov Models and Artificial Neural Networks for Speech and Speaker Recognition

  Date: September 7, 1998

- **School:** Faculté Polytechnique of Mons, Belgique

  Subject: PhD thesis committee

  Expert: Prof. Hervé Bourlard

  Candidate: Olivier Deroo

Title: Modèles dépendants du contexte et méthodes de fusion de données appliqués à la reconnaissance de la parole par modèles hybrides HMM/MLP

Date: September 7, 1998

- **School:** École Supérieure d'Informatique de Sierre (ESIS), Valais

  Subject: Diploma work exams

  Expert: Dominique Genoud

  Date: December 17, 1998

5 Other Scientific Activities

5.1 Editorship

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

5.2 Scientific Committees Membership

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

Function: Member of the Administration Committee

Conference: European Association for Signal Processing (EURASIP)

- **Name:** Dr. Juergen Luettin

  Function: Member of the Scientific Committee

  Conference: International Conference on Human Interfaces In Control Rooms, Cockpits and Command Centres, Bath, UK

- **Name:** Dr. Eddy Mayoraz

  Function: Member of the Scientific Committee

  Conference: European Symposium of Artificial Neural Networks (ESANN)

- **Name:** Dr. Eddy Mayoraz

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

  Conference: 5th International Symposium on Artificial Intelligence and Mathematics

- **Name:** Dr. Georg Thimm

  Function: Current Events Editor

  Journal: Neurocomputing

5.3 Organization of Conference

- **Title:** XXII^{emes} Journées D'Étude sur la Parole (JEP'98)

  Location: CERM, Martigny, Valais, Switzerland

  Date: June 15 – 19, 1998

  Conference Board: *General Chair*: Hervé Bourlard (IDIAP, Switzerland); *Secretary*: Chantal Pillet (Switzerland); *Organizers*: Olivier Bornet (IDIAP, Switzerland), Gilles Caloz (IDIAP, Switzerland), Frank Formaz (IDIAP, Switzerland), Dominique Genoud (IDIAP, Switzerland), Cédric Jaboulet (UBILAB, Switzerland), Sacha Krstulovic (IDIAP, Switzerland), Johnny Mariethoz (IDIAP, Switzerland);

  Organisers: IDIAP and GFCP (Groupe Francophone de la Communication Parlée)

  Sponsors: SFA (Société Française d'Acoustique), ESCA (European Speech Communication Association), Swisscom, SNSF (Swiss National Science Foundation), the City of Martigny, SUN Microsystems et la Société Académique du Valais (SAV)

  Scientific Committee: M. Adda-Decker (France), R. André-Obrecht (France), F. Bimbot (France), J.-F. Bonastre (France), H. Bourlard (Switzerland), J.-L. Cochard (Switzerland), P. Dupont (France), P. Deléglise (France), J.-M. Hombert (France), Y. Laprie (France), R.K. Moore (UK), C. Montacié (France), P. Perrier (France), J. Schoentgen (Belgium), R. Sock (France), B. Teston (France), J. Vaissière (France)

5.4 Short term visits

- **Location:** Institut de la Communication Parlée, ICP, Institut National Polytechnique, Grenoble, France

  Visitor: Hervé Glotin

  Date: 30% of the year.

- **Location:** Institut de la Communication Parlée (ICP), INPG, Grenoble, France

Visitor: Dr. Andrew Morris

Date: March 20, 1998

- **Location:** Faculté Polytechnique Mons (FPMS), Belgium

  Visitor: Astrid Hagen

  Date: March 30 – April 3, 1998

- **Location:** Speech and Hearing research group, Sheffield University Department of Computer Science, UK

  Visitor: Dr. Andrew Morris

  Date: 24 July and 21 December, 1998

- **Location:** Rutgers Center for Operations Research (RUTCOR), NJ, USA

  Visitor: Dr. E. Mayoraz

  Date: July 27 – August 31, 1998

- **Location:** International Computer Science Institute, Berkeley, CA, USA

  Visitor: Prof. H. Bourlard

  Date: July 15 – August 15, 1998

6 Events and Presentations

6.1 Scientific Presentations

- **Event:** Interdisziplinäres Kolleg, Spring Scholl (IK'98), Günne am Möhnessee, Germany, March 9-11, 1998

  Invited lecturer: Prof. H. Bourlard

  Title: Neural Networks and Conventional Algorithms

- **Event:** Interdisziplinäres Kolleg, Spring Scholl (IK'98), Günne am Möhnessee, Germany, March 7–14, 1998

  Speaker: Prof. H. Bourlard [p-Bou98]

- **Event:** European Conference on Machine Learning (ECML'98), Chemnitz, Germany,

  Speaker: Miguel Moreira [p-MM98]

- **Event:** The International Conference of Acoustics, Speech and Signal Processing, Seattle, WA, USA, May 12-15, 1998

  Speaker: Prof. H. Bourlard [p-GMM98]

- **Event:** The International Conference of Acoustics, Speech and Signal Processing, Seattle, WA, USA, May 12-15, 1998

  Speaker: Dr. Andrew Morris

  Title: Some solutions to the missing feature problem in data classification, with application to noise robust ASR

- **Event:** Neurosciences et Sciences pour l'Ingénieur, NSI'98, Munster, France, May 11-15, 1998

  Speaker: Hervé Glotin [p-GTBB98b]

- **Event:** Journées Etudes Parole, JEP'98, Martigny, June 15-19, 1998

  Speaker: Hervé Glotin [p-GTBB98a]

- **Event:** 5th European Conference on Computer Vision, Freiburg, Germany, June 2-6, 1998

  Speaker: Juergen Luettin [p-LD98]

- **Event:** Speech and Dialog International Workshop, TSD'98, Brno, Czech Republic, September 23-26, 1998

  Speaker: Hervé Glotin [p-HBTB98]

- **Event:** Workshop on Text, Speech and Dialog (TSD'98), Brno, Czech Republic, September 23–26, 1998

  Speaker: Giulia Bernardis [p-BB98a]

- **Event:** International Conference on Speech and Language Processing, ICSLP'98, Sidney, Australia, December 1-4, 1998

  Speaker: Hervé Glotin [p-BGTB98]

- **Event:** International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, Nov. 30 – Dec. 4, 1998

  Speaker: Prof. H. Bourlard [p-BB98b]

6.2 Regional Presentations

In 1998, IDIAP's activities were presented to the broad public at the regional fairs: La Foire du Valais, Martigny, October 8–9, 1998.

On October 23, 1998, the Federal Chancellerie visited IDIAP. After a welcome speech from President Pierre Crittin, Prof. Hervé Bourlard presented an overview of the main activities of IDIAP highlighted by some demonstrations of the different technologies developed in the Institute.

7 Publications (1997 and 1998)

7.1 Books and Book Chapters

[b-BBD+99] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. *Traitement de la Parole.* Presses Polytechniques Universitaires Romandes, 1999.

[b-BM98] H. Bourlard and N. Morgan. *Survey of the State of the Art in Human Language Technology*, chapter Connectionist Techniques, pages 356–361. Cambridge University Press, 1998.

[b-Fie97] *CRC Comprehensive Dictionary of Electrical Engineering.* CRC Press, Boca Raton, Florida, 1997. Contributing Author: E. Fiesler.

[b-Lue99] J. Luettin. Speech reading. In J. Noyes and M. Cooke, editors, *Modern Interface Technology: The Leading Edge*, pages 97–121. Research Studies Press Ltd., 1999.

[b-MBH98] N. Morgan, H. Bourlard, and H. Hermansky. Automatic speech recognition: an auditory perspective. In S. Greenberg, W. Ainsworth, A. Popper, and R. Fay, editors, *Speech Processing in the Auditory System.* Springer Verlag, New York, 1998. IDIAP-RR 98-17.

[b-MF97] P. D. Moerland and E. Fiesler. Neural network adaptations to hardware implementations. In E. Fiesler and R. Beale, editors, *Handbook of Neural Computation*, pages E1.2:1–13. Institute of Physics Publishing and Oxford University Publishing, New York, 1997. IDIAP-RR 97-17.

[b-Sax97] I. Saxena. Ellipsometry. In P. K. Rastogi, editor, *Optical Metrology.* Artech House, 1997.

7.2 Articles in International Journals

[a-BKT97] M. Bader, W. E. Klee, and G. Thimm. The 3-regular nets with 4 and 6 vertices per unit cell. *Zeitschrift für Kristallographie*, 212:553–558, 1997.

[a-DBB+97] Benoît Duc, Elizabeth Saers Bigün, Josef Bigün, Gilbert Maître, and Stefan Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):835–843, 1997.

[a-JLGW97] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):853–858, 1997. IDIAP-RR 97-13.

[a-LT97] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997. IDIAP-RR 97-12.

[a-May99] Eddy Mayoraz. On the complexity of recognizing regions computable by two-layered perceptrons. *Annals Mathematics and Artificial Intelligence*, 1999. To appear.

[a-MFS98] P. D. Moerland, E. Fiesler, and I. Saxena. Discrete all-positive multilayer perceptrons for optical implementation. *Optical Engineering*, 37(4):1305–1315, April 1998. (IDIAP-RR 97-02).

[a-TF97a] Georg Thimm and Emile Fiesler. High order and multilayer perceptron initialization. *IEEE Transactions on Neural Networks*, 8(2), 1997.

[a-TF97b] Georg Thimm and Emile Fiesler. Two neural network construction methods. *Neural Processing Letters*, 6(1), 1997.

[a-Thi97] Georg Thimm. Calendar of meetings (several issues). *Neurocomputing*, 1997. published since 1995.

[a-TK97] G. Thimm and W. E. Klee. Zeolite cycle sequences. *Zeolites*, 19:422–424, 1997.

7.3 Articles in Conference Proceedings

[p-BB98a] Giulia Bernardis and Hervé Bourlard. Confidence measures in hybrid HMM/ANN speech recognition. In *Proceedings of Workshop on Text, Speech and Dialog (TSD'98) Brno, Czech Republic*, pages 159–164, September 1998.

[p-BB98b] Giulia Bernardis and Hervé Bourlard. Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'98) Sydney, Australia*, pages 775–778, 1998. IDIAP-RR 98-11.

[p-BD97] H. Bourlard and S. Dupont. Subband-based speech recognition. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1251–1254, April 1997.

[p-BG97] F. Bimbot and D. Genoud. Likelihood ratio adjustment for the compensation of model mismatch in speaker verification. In *Eurospeech 97*, 1997. IDIAP-RR 97-05.

[p-BGTB98] F. Berthommier, H. Glotin, E. Tessier, and H. Bourlard. Interfacing of CASA and partial recognition based on a multistream technique. In *ICSLP'98*, volume 4, pages 1415–1419. Sidney, 1998.

[p-BHJ+97] F. Bimbot, H-P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J-B. Pierrot. Speaker verification in the telephone network : Research activities in the CAVE project. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, 1997.

[p-BM97] Hervé Bourlard and Nelson Morgan. Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. In *International School on Neural Nets: Adaptive Processing of Temporal Information*. Springer Verlag, 1997. To appear.

[p-Bou97] Hervé Bourlard. State-of-the-art and recent progress in hybrid HMM/ANN speech recognition. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'97)*, number 1327 in Lecture Notes in Computer Science, pages 875–884. Springer-Verlag, 1997.

[p-Bou98] Hervé Bourlard. Connectionist speech recognition. In *Proceedings of IK'98, Interdisziplinäres Kolleg, Spring Scholl, Günne am Möhnessee, Germany, March 7–14*, pages 61–89, 1998.

[p-BY99] S. Ben-Yacoub. Multi-modal data fusion for person authentication using SVM. In *Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99)*, pages 25–30, 1999.

[p-BYFL99] S. Ben-Yacoub, B. Fasel, and J. Luettin. Fast face detection using MLP and FFT. In *Proc. Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, pages 31–36, 1999.

[p-BYJJ+99] S. Ben-Yacoub, J.Luettin, K. Jonsson, J. Matas, and J. Kittler. Audio-visual person verification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 1999, Fort Collins, USA*, 1999. IDIAP-RR 98-18.

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

[p-DB97] S. Dupont and H. Bourlard. Using multiple time scales in a multi-stream speech recognition system. In *EUROSPEECH'97*, pages 3–6, September 1997.

[p-DBD+97] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite. Hybrid HMM/ANN systems for training independent tasks: Experiments on 'phonebook' and related improvements. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1767–1770, April 1997.

[p-DBR97] Stéphane Dupont, Hervé Bourlard, and Christophe Ris. Robust speech recognition based on multi-stream features. In *Proc. of the ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 95–98, April 1997. IDIAP-RR 97-01.

[p-DL98] S. Dupont and J. Luettin. Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database. In *Proc. 5th Int. Conf. on Spoken Language Processing*, volume 4, pages 1283–1286, 1998.

[p-DMFB97] Benoît Duc, Gilbert Maître, Stefan Fischer, and Josef Bigün. Person authentication by fusing face and speech information. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science, pages 311–318. Springer Verlag, 1997.

[p-ea99] G. Richard et al. Multi modal verification for teleservices and security applications. In *IEEE International Conference on Multimedia Computing and Systems*, 1999.

[p-FB97] V. Fontaine and H. Bourlard. Speaker-dependent speech recognition based on phone-like unit model – application to voice dialing. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1527–1530, April 1997.

[p-FHC+98] Bimbot F., Hutter H.P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.B. An overview of the cave project research activities in speaker verification. In *Reconnaissance du locuteur et ses applications commerciales et criminalistiques*, 1998. To appear.

[p-FM97] E. Fiesler and M. Maignan. A connectionist system for two-dimensional representation of multivariate location data. In *Proceedings of the Fifth International Workshop on Artificial Intelligence for High Energy Physics*, Amsterdam, The Netherlands, 1997. AIHENP, Elsevier Science.

[p-GMM98] Dominique Genoud, Miguel Moreira, and Eddy Mayoraz. Text dependent speaker verification using binary classifiers. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing — ICASSP'98*, volume I, pages 129–132. IEEE, IEEE, May 1998. IDIAP-RR 97-08.

[p-GTBB98a] H. Glotin, E. Tessier, H. Bourlard, and F. Berthommier. Reconnaissance multi-bandes de la parole bruitée par couplage entre les niveaux primitifs et d'identification. In *Journées Etude Parole - Martigny*, Juin 1998.

[p-GTBB98b] H. Glotin, E. Tessier, H. Bourlard, and F. Berthommier. Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes. In *Neurosciences et Sciences de l'Ingénieur'98 - Munster, CNRS*, Mai 1998.

[p-HBTB98] H.Glotin, F. Berthommier, E. Tessier, and H. Bourlard. Interfacing of CASA and multistream recognition. In *TSD'98-Text, Speech and Dialog International Workshop*. BRNO-Czech Republic, Sept 1998.

[p-HRBR97] J. Hennebert, C. Ris, H. Bourlard, and S. Renals. Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. In *EUROSPEECH'97*, pages 1951–1954, September 1997.

[p-JJJ+98] Pierrot J.B., Lindberg J., Koolwaaij J., Hutter H.P., Genoud D., Blomberg M., and Bimbot F. A comparison of a priori threshold setting procedures for speaker verification in the cave project. In *ICASSP 98*, 1998. To appear.

[p-JLGW97a] Pierre Jourlin, Juergen Luettin, Dominique Genoud, and Hubert Wassner. Acoustic-labial speaker verification. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science, pages 319–326. Springer Verlag, 1997. IDIAP-RR 97-13.

[p-JLGW97b] P Jourlin, J. Luettin, D. Genoud, and H. Wassner. Integrating acoustic and labial information for speaker identification and verification. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1603–1606, 1997.

[p-LD98] J. Luettin and S. Dupont. Continuous audio-visual speech recognition. In *Proc. 5th European Conference on Computer Vision*, volume II of *Lecture Notes in Computer Science*, pages 657–673. Springer Verlag, 1998.

[p-Lue97] J. Luettin. Towards speaker independent continuous speechreading. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1991–1994, 1997.

[p-May97] Eddy Mayoraz. On the complexity of recognizing iterated differences of polyhedra. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'97)*, number 1327 in Lecture Notes in Computer Science, pages 475–480. Springer-Verlag, 1997. IDIAP-RR 97-10.

[p-MC99] Chafik Mokbel and Olivier Collin. Incremental enrollment of speech recognizers. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing — ICASSP'99*. IEEE, IEEE, 1999.

[p-MM97] Eddy Mayoraz and Miguel Moreira. On the decomposition of polychotomies into dichotomies. In *Proceedings of The Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997. IDIAP-RR 96-08.

[p-MM98] Miguel Moreira and Eddy Mayoraz. Improved pairwise coupling classification with correcting classifiers. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Artificial Intelligence*, pages 160–171. Springer, April 1998. IDIAP-RR 97-09.

[p-MMK+99] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99)*, 1999.

[p-Moe97] Perry Moerland. Mixtures of experts estimate a posteriori probabilities. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'97)*, number 1327 in Lecture Notes in Computer Science, pages 499–504, Berlin, 1997. Springer-Verlag.

[p-PDHGC96] D. Petrovska-Delacretaz, J. Hennebert, D. Genoud, and G. Chollet. Semi-automatic hmm-based annotation of the polycost database. In *Application of speaker recognition techniques in telephony*. COST250, 1996.

[p-PHMG98] D. Petrovska, J. Hennebert, H. Melin, and D. Genoud. Polycost: a telephone-speech database for speaker recognition. In *Reconnaissance du locuteur et ses applications commerciales et criminalistiques*, 1998. To appear.

[p-SMF+97] I. Saxena, P. Moerland, E. Fiesler, A. R. Pourzand, and N. Collings. An optical thresholding perceptron. In *Proceedings of the Workshop on Optics and Computer Science*, 1997. IDIAP-RR 97-16.

[p-SMFP97] I. Saxena, P. Moerland, E. Fiesler, and A. Pourzand. Handwritten digit recognition with binary optical perceptron. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'97)*, number 1327 in Lecture Notes in Computer Science, pages 1253–1258, Berlin, 1997. Springer-Verlag. IDIAP-RR 97-15.

[p-TEFK97] M. Tajine, D. Elizondo, E. Fiesler, and J. Korczak. Adapting the 2-class recursive deterministic perceptron neural network to m classes. In *Proceedings of the International Conference on Neural Networks*. IEEE, IEEE, 1997.

[p-TL98] G. Thimm and J. Luettin. Illumination-robust pattern matching using distorted color histograms. In *Lecture Notes in Computer Science (5th Open German-Russian Workshop on Pattern Recognition and Image Understanding)*. Springer Verlag, September 21 - 25, 1998. To appear.

[p-VMM98] Patrick Verlinde, Gilbert Maître, and Eddy Mayoraz. Decision fusion using a multi-linear classifier. In *1st International Conference on Multisource-Multisensor Data Fusion*, July 1998. To appear.

7.4 IDIAP Research Reports

[r-Alp98] Ethem Alpaydin. Combined 5x2cv f-test for comparing supervised classification learning algorithms. IDIAP-RR 4, IDIAP, 1998. Submitted for publication.

[r-AM98] Ethem Alpaydin and Eddy Mayoraz. Combining linear dichomotizers to construct nonlinear polychotomizers. IDIAP-RR 5, IDIAP, 1998.

[r-Beu97] Jean-Luc Beuchat. Reconnaissance de caractères manuscrits à l'aide de réseaux neuromimétiques. IDIAP-RR 18, IDIAP, 1997.

[r-BM98] Hervé Bourlard and Nelson Morgan. Speaker verification: A quick overview. IDIAP-RR 12, IDIAP, 1998.

[r-Bou98] Hervé Bourlard. Introduction à la reconnaissance de la parole et du locuteur. IDIAP-RR 13, IDIAP, 1998. Chapitre du livre Traitement de la Parole, à paraître aux Presses Polytechniques Universitaires Romandes.

[r-BY97] Souheil Ben-Yacoub. Fast object detection using MLP and FFT. IDIAP-RR 11, IDIAP, 1997.

[r-DL97] S. Dupont and J. Luettin. Using the multi-stream approach for continuous audio-visual speech recognition. IDIAP-RR 14, IDIAP, 1997.

[r-HMB98] A. Hagen, A. Morris, and H. Bourlard. Subband-based speech recognition in noisy conditions: The full combination approach. IDIAP-RR 15, IDIAP, 1998.

[r-Krs97] Sacha Krstulović. Investigation of a possible process identity between DRM and linear filtering. IDIAP-RR 19, IDIAP, 1997.

[r-Krs98] Sacha Krstulović. Acoustico-articulatory inversion of the DRM model through inverse filtering. IDIAP-RR 16, IDIAP, 1998.

[r-MA98] Eddy Mayoraz and Ethem Alpaydin. Support vector machine for multiclass classification. IDIAP-RR 6, IDIAP, 1998. Submitted for publication.

[r-Moe98] Perry Moerland. Localized mixtures of experts. IDIAP-RR 14, IDIAP, 1998.

[r-TBYL98] G. Thimm, S. Ben-Yacoub, and J. Luettin. Evaluating the complexity of databases for person identification and verification. IDIAP-RR 10, IDIAP, August 1998.

[r-TF97a] G. Thimm and E. Fiesler. Optimal setting of weights, learning rate, and gain. IDIAP-RR 04, IDIAP, 1997.

[r-TF97b] G. Thimm and E. Fiesler. Pruning of neural networks. IDIAP-RR 03, IDIAP, 1997.

[r-TL98a] G. Thimm and J. Luettin. Illumination-robust pattern matching using distorted color histograms. IDIAP-RR 9, IDIAP, Rue de Simplon 4, CP 592, CH-1920 Martigny, Switzerland. Email: Thimm@idiap.ch, June 1998.

[r-TL98b] G. Thimm and J. Luettin. Optimal parameterization of point distribution models. IDIAP-RR 01, IDIAP, 1998.

[r-VMM97] Patrick Verlinde, Gilbert Maître, and Eddy Mayoraz. Decision fusion in a multi-modal identity verification system using a multi-linear classifier. IDIAP-RR 6, IDIAP, 1997.

7.5 IDIAP Communications

[c-ACB97] Johan M. Andersen, Gilles Caloz, and Hervé Bourlard. Swisscom "AVIS" project (no. 392) advanced vocal interfaces services. IDIAP-COM 6, IDIAP, 1997.

[c-And98] Johan M. Andersen. Baseline system for hybrid speech recognition on french (experiments on bref). IDIAP-COM 07, IDIAP, 1998.

[c-BCF+97] Hervé Bourlard, Jean-Luc Cochard, Emile Fiesler, Gilbert Maitre, and Eddy Mayoraz. Activity report 1996. IDIAP-Com 1, IDIAP, 1997.

[c-Fas98] Beat Fasel. Fast multi-scale face detection. IDIAP-COM 4, IDIAP, 1998.

[c-GC97] D. Genoud and G. Caloz. 1997 NIST evaluation: Text independent speaker detection (verification). IDIAP-Com 3, IDIAP, 1997.

[c-LM97] Tomas Lundin and Perry Moerland. Quantization and pruning of multilayer perceptrons: Towards compact neural networks. IDIAP-Com 2, IDIAP, March 1997.

[c-LM98] J. Luettin and G. Maître. Evaluation protocol for the extended M2VTS database (XM2VTSDB). IDIAP-COM 05, IDIAP, 1998.

[c-Moe97] Perry Moerland. Some methods for training mixtures of experts. IDIAP-Com 5, IDIAP, November 1997.

[c-Sch97] Michael Schmal. Speaker verification by pairwise coupling. IDIAP-Com 7, IDIAP, October 1997.

[c-Van97] Samuel Vannay. Réalisation d'un majordome vocal. IDIAP-Com 4, EPFL / IDIAP, 1997.

7.6 Other Documents

[o-Gen99] Dominique Genoud. *Reconnaissance et Transformation de Locuteurs.* PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, January 1999.

[o-Lue97] Juergen Luettin. *Visual Speech and Speaker Recognition.* PhD thesis, University of Sheffield, 1997.

[o-Thi97] Georg Thimm. *Optimization of high order perceptrons.* PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, June 1997. Dissertation number 1633.