PRINCIPAL 2003 SPONSORS

Intel
Qualcomm
IM2 National Centre of Competence in Research, Switzerland
Förderverein (German industrial organization)
German Ministry of Research and Technology (via U. Munich)
Spanish Ministry of Science and Technology (MCYT)
Finnish National Technology Agency (TEKES)
National Science Foundation
DARPA

AFFILIATED 2003 SPONSORS

European Media Lab
European Union (via Sheffield University)
HP
Infineon
Microsoft

CORPORATE OFFICERS

Prof. Nelson Morgan (President and Institute Director)
Prof. Scott Shenker (Vice President)
Dr. Jordan Cohen (Secretary and Treasurer)

BOARD OF TRUSTEES

Dr. Charlie Bass, The Bass Trust
Prof. Hervé Bourlard, IDIAP and EPFL, Switzerland
Vice Chancellor Beth Burnside, UC Berkeley
Dr. Jordan Cohen, VoiceSignal Inc.
Dr. Adele Goldberg, Neometron
Dr. Greg Heinzinger, Qualcomm
Mr. Clifford Higgerson, Convventures (Chairman)
Prof. Richard Karp, ICSI and UC Berkeley
Prof. Pedro Lizzano, Telefonica
Prof. Jitendra Malik, UC Berkeley
Prof. Nelson Morgan, ICSI (Director) and UC Berkeley
Dr. David Nagel, PalmSource
Dr. Ilpo Reitman, TEKES
Prof. Shankar Sastry, UC Berkeley
Prof. Scott Shenker, ICSI and UC Berkeley
Prof. Wolfgang Wahlster, DFKI GmbH
### 2003 Visitors in Sponsored International Programs

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Group</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrei Gurtov</td>
<td>Finland</td>
<td>Networks</td>
<td>TEKES</td>
</tr>
<tr>
<td>Pekka Himanen</td>
<td>Finland</td>
<td>BCIS</td>
<td>TEKES</td>
</tr>
<tr>
<td>Petri Kiukkonen</td>
<td>Finland</td>
<td>AI</td>
<td>TEKES</td>
</tr>
<tr>
<td>Konsta Koppinen</td>
<td>Finland</td>
<td>Speech</td>
<td>TEKES</td>
</tr>
<tr>
<td>Vesa-Matti Lahti</td>
<td>Finland</td>
<td>BCIS</td>
<td>TEKES</td>
</tr>
<tr>
<td>Olli Leppanen</td>
<td>Finland</td>
<td>BCIS</td>
<td>TEKES</td>
</tr>
<tr>
<td>Kim Lopperi</td>
<td>Finland</td>
<td>Haas</td>
<td>TEKES</td>
</tr>
<tr>
<td>Maarit Makinen</td>
<td>Finland</td>
<td>BCIS</td>
<td>TEKES</td>
</tr>
<tr>
<td>Tuomo Pirinen</td>
<td>Finland</td>
<td>Speech</td>
<td>TEKES</td>
</tr>
<tr>
<td>Jusso Rantala</td>
<td>Finland</td>
<td>Haas</td>
<td>TEKES</td>
</tr>
<tr>
<td>Sven Behnke</td>
<td>Germany</td>
<td>Speech</td>
<td>DAAD</td>
</tr>
<tr>
<td>Jens Gramm</td>
<td>Germany</td>
<td>Algorithms</td>
<td>DAAD</td>
</tr>
<tr>
<td>Marek Musial</td>
<td>Germany</td>
<td>AI</td>
<td>DAAD</td>
</tr>
<tr>
<td>Till Nierhoff</td>
<td>Germany</td>
<td>Algorithms</td>
<td>DAAD</td>
</tr>
<tr>
<td>Petra Steiner</td>
<td>Germany</td>
<td>AI</td>
<td>DAAD</td>
</tr>
<tr>
<td>Till Tantau</td>
<td>Germany</td>
<td>Algorithms</td>
<td>DAAD</td>
</tr>
<tr>
<td>Klaus Wehrle</td>
<td>Germany</td>
<td>Networks</td>
<td>DAAD</td>
</tr>
<tr>
<td>Oliver Wendt</td>
<td>Germany</td>
<td>Algorithms</td>
<td>FV</td>
</tr>
<tr>
<td>Britta Wrede</td>
<td>Germany</td>
<td>Speech</td>
<td>DAAD</td>
</tr>
<tr>
<td>Fernando de Arriaga</td>
<td>Spain</td>
<td>AI</td>
<td>MCYT</td>
</tr>
<tr>
<td>Javier Cardona</td>
<td>Spain</td>
<td>Networks</td>
<td>MCYT</td>
</tr>
<tr>
<td>José Gonzalez</td>
<td>Spain</td>
<td>Networks</td>
<td>MCYT</td>
</tr>
<tr>
<td>Javier Macias</td>
<td>Spain</td>
<td>Speech</td>
<td>MCYT</td>
</tr>
<tr>
<td>Sira Palazuelos</td>
<td>Spain</td>
<td>AI</td>
<td>MCYT</td>
</tr>
<tr>
<td>Carlos Subirats</td>
<td>Spain</td>
<td>AI</td>
<td>MCYT</td>
</tr>
<tr>
<td>Jitendra Ajmera</td>
<td>Switzerland</td>
<td>Speech</td>
<td>IM2</td>
</tr>
<tr>
<td>Sebastien Coquoz</td>
<td>Switzerland</td>
<td>Speech</td>
<td>IM2</td>
</tr>
<tr>
<td>Micha Hersch</td>
<td>Switzerland</td>
<td>Speech</td>
<td>IM2</td>
</tr>
</tbody>
</table>

DAAD: Deutscher Akademischer Austauschdienst       TEKES: Finnish National Technology Agency
MCYT: Ministerio de Ciencia y Tecnología, Estado de Política Científica y Tecnología
IM2: Interactive Multimodal Information Management, National Centre of Competence in Research, Switzerland
FV: Förderverein, Germany
Contents

I  INSTITUTE OVERVIEW  1

1  Institute Sponsorship for 2003  1

2  Institutional Structure of ICSI  2
  2.1  Management and Administration  2
  2.2  Research  3
    2.2.1  Senior Research Staff  3

II  RESEARCH GROUP REPORTS  5

1  Research Group Highlights  5
  1.1  Networking  5
  1.2  Algorithms  5
  1.3  AI  6
  1.4  Speech  6
  1.5  BCIS highlights  6

2  Networking  7
  2.1  Internet Architecture  7
    2.1.1  Host-based Congestion Control  7
    2.1.2  Router-based Congestion Control  8
    2.1.3  Routing  9
    2.1.4  Novel Approaches  10
  2.2  Measurements and Modeling  11
  2.3  Peer-to-Peer Systems  12
    2.3.1  Distributed Hash Tables  12
    2.3.2  Other P2P Systems  13
  2.4  Security, Malware, and Intrusion Detection  14
  2.5  Extensible Open Router Platform  17
  2.6  Sensornets  17
  2.7  Internet Community Activities  18

3  Algorithms  23
  3.1  Overview  23
  3.2  Computational Biology  23
    3.2.1  Systems Biology  23
    3.2.2  Genome Rearrangements  25
    3.2.3  SNP Genotyping  25
    3.2.4  Gene Expression Analysis  26
    3.2.5  Transcriptional Regulation  26
    3.2.6  Haplotyping  28
    3.2.7  Phylogeny  29
3.3 Risk, Reward, Reinforcement and Computational Finance .......... 30
3.4 Combinatorial Algorithms ........................................... 31
  3.4.1 Edge Modification Problems ................................. 31
  3.4.2 The Stacker Crane Problem .................................... 34
  3.4.3 Arc-Annotated Sequences ..................................... 34
  3.4.4 Fixed-parameter tractability and approximation .......... 34
3.5 Networking ......................................................... 35
3.6 Computational Complexity ......................................... 35
  3.6.1 Reducibility to Sets with Low Information Content ........ 35
  3.6.2 CVerbose Languages ........................................... 36
  3.6.3 Reachability in Digraphs with Low Independence Number .... 36
3.7 Research on Combinatorial Games ................................ 36
  3.7.1 Quantitative Go ............................................... 37
  3.7.2 Cg suite ...................................................... 37

4 Artificial Intelligence and its Applications .......................... 44
  4.1 The Neural Theory of Language, J. Feldman and S. Narayanan .... 44
     4.1.1 Semantic Web Services ....................................... 47
     4.1.2 Applications of FrameNet .................................... 47
     4.1.3 CPRM: A Scalable and Expressive Model of Actions and Events . 48
  4.2 FrameNet, C. Fillmore ............................................ 49
     4.2.1 Collaborations Initiated or Continued ....................... 51
  4.3 Connectionist Models of Brain Function, L. Shastri and S. Narayanan 52
     4.3.1 Modeling cultural differences ................................ 52
     4.3.2 Cognitive control ............................................ 53
     4.3.3 Learning structured representations ......................... 53
     4.3.4 A biologically realistic model of episodic memory .......... 53
     4.3.5 The Role of Cortico-SubCortical Loops in Planning and Working Memory: A Computational Model ....................... 53

5 Speech Processing .................................................. 57
  5.1 EARS .............................................................. 57
     5.1.1 Rich Transcription ......................................... 58
     5.1.2 Novel Approaches ............................................ 61
  5.2 Speech from Meetings ............................................. 64
     5.2.1 Corpus Development ......................................... 64
     5.2.2 Dialog Acts .................................................. 64
     5.2.3 Detection of Meeting Hotspots .............................. 65
     5.2.4 Discourse markers of meeting content and social structure .. 66
  5.3 Spoken Language Systems ........................................ 67
  5.4 Speaker Recognition: Modeling Idiosyncrasies in Speaking Behavior .. 68
  5.5 Computer architectures for speech processing .................... 69
  5.6 Physiologically motivated speech processing ..................... 69
  5.7 ICT4B ............................................................. 70

6 Berkeley Center for the Information Society .......................... 73
7 Robust Video Compression based on Distributed Source Coding techniques

7.1 Summary of Results using Multilevel Coset Codes for PRISM .................................. 75
7.2 Improved Resilience to Drift through the use of a Side-Channel ................................. 76
Part I

INSTITUTE OVERVIEW

The International Computer Science Institute (ICSI) is an independent, nonprofit basic research institute affiliated with the University of California campus in Berkeley, California. Its establishment was motivated by recognition of the need for an international fundamental research facility in the field of computer science. ICSI was started in 1986 and inaugurated in 1988 as a joint project of the Computer Science Division of UC Berkeley and the GMD, the Research Center for Information Technology GmbH in Germany. Since then, Institute collaborations within the university have broadened (for instance, with the Electrical Engineering Division, as well as other departments such as Linguistics). In addition, Institute support has expanded to include a range of international collaborations, US Federal grants, and direct industrial sponsorship. Throughout these changes, the Institute has maintained its commitment to a pre-competitive, open research program. The goal of the Institute continues to be the creation of synergy between leading academic and industrial research in an international environment through excellence in fundamental research in computer science and engineering.

The particular areas of concentration have varied over time, but are always chosen for their fundamental importance and their compatibility with the strengths of the Institute and affiliated UC Berkeley staff. ICSI currently has significant efforts in five primary research areas: Internet research, including Internet architecture, related theoretical questions, and network security; theoretical computer science, including applications to the modeling of both biological and internet-related phenomena; artificial intelligence, particularly for applications to natural language understanding, but also for biological modeling; natural speech processing, including both speech and speaker recognition; and the study of the social impact of information technology.

The Institute occupies a 28,000 square foot research facility at 1947 Center Street, just off the central UC campus in downtown Berkeley. Administrative staff provide support for researchers: housing, visas, computational requirements, grants administration, etc. There are approximately eighty scientists in residence at ICSI including permanent staff, postdoctoral Fellows, visitors, affiliated faculty, and students. Senior investigators are listed at the end of this overview, along with their current interests.

1 Institute Sponsorship for 2003

As noted earlier, ICSI is sponsored by a range of US Federal, international, and industrial sources. The figure below gives the relative distribution of funding among these different sponsoring mechanisms.

US Federal funding comes from a range of grants to support research Institute-wide. Most of this funding comes from the National Science Foundation and DARPA. International support in 2003 came from government and industrial programs in Germany, the Ministry of Science and Technology in Spain, the National Technology Association of Finland, and the Swiss National Science Foundation (through the Swiss Research Network IM2). Additional support comes from the European Media Lab and the European Union. Industrial support in 2003 was
primarily provided by Intel and Qualcomm, with additional sponsorship from Infineon, HP, and Microsoft. We also benefited significantly in 2003 from prior funding from AT&T.

ICSI’s budget increased in 2003 (about 11% over 2002), to roughly $8.9M for the year.

2 Institutional Structure of ICSI

ICSI is a nonprofit California corporation with an organizational structure and bylaws consistent with that classification and with the institutional goals described in this document. In the following sections we describe the two major components of the Institute’s structure: the Administrative and Research organizations.

2.1 Management and Administration

The corporate responsibility for ICSI is ultimately vested in the person of the Board of Trustees, listed in the first part of this document. Ongoing operation of the Institute is the responsibility of Corporation Officers, namely the President, Vice President, and Secretary-Treasurer. The
President also serves as the Director of the Institute, and as such, takes responsibility for day-to-day Institute operations.

Internal support functions are provided by three departments: Computer Systems, Finance, and Administrative Operations. Computer Systems provides support for the ICSI computational infrastructure, and is led by the Systems Manager. Finance is responsible for payroll, grants administration, benefits, human resources, and generally all Institute financial matters; it is led by the Controller. All other support activities come under the general heading of Administrative Operations, and are supervised by the Operations Manager; these activities include office assignments, housing, visas, grant proposal administration, and support functions for ongoing operations and special events.

2.2 Research

Research at ICSI is overwhelmingly investigator-driven, and themes change over time as they would in an academic department. Consequently, the interests of the senior research staff are a more reliable guide to future research directions than any particular structural formalism. Nonetheless, ICSI research has been organized into Groups: the Networking Group (Internet research), the Algorithms Group, the AI Group, the Speech Group, and the Berkeley Center for the Information Society. Consistent with this organization, the bulk of this report is organized along these lines, with one sub-report for each of the five groups. There is also an additional activity in video encoding that does not fall neatly into any of the groups, and which is described at the report’s end.

Across all of these activities, there is a theme: scientific studies based on the growing ubiquity of connected computational devices. In the case of Networking studies, the focus is on the Internet; in the case of Speech and AI, it is on the interfaces to the distributed computational devices; and in the case of our new Center, it is on the social impact of these technologies. Finally, the Algorithms group continues to develop methods that are employed in a range of computational problems.

2.2.1 Senior Research Staff

The previous paragraphs briefly described the clustering of ICSI research into major research themes and working groups. Future work could be extended to new major areas based on strategic Institutional decisions and on the availability of funding to support the development of the necessary infrastructure. At any given time, though, ICSI research is best seen as a set of topics that are consistent with the interests of the Research Staff. In this section, we give the names of current (March 2004) senior research staff members at ICSI, along with a brief description of their current interests and the Research Group that the researcher is most closely associated with. This is probably the best snapshot of research directions for potential visitors or collaborators. Not shown here are the range of postdoctoral Fellows, visitors, and graduate students who are also key contributors to the intellectual environment at ICSI.

Mark Allman (Networking): congestion control, network measurement, network dynamics, transport protocols and network security.

Jerome Feldman (AI): neural plausible (connectionist) models of language, perception and learning and their applications.
Charles Fillmore (AI): building a lexical database for English (and the basis for multilingual expansion) which records facts about semantic and syntactic combinatorial possibilities for lexical items, capable of functioning in various applications: word sense disambiguation, computer-assisted translation, information extraction, etc.

Sally Floyd (Networking): congestion control, transport protocols, queue management, and network simulation.

Atanu Ghosh (Networking): extensible open source routing, active networks, protocols, multimedia, and operating systems.


Nelson Morgan (Speech): signal processing and pattern recognition, particularly for speech and biomedical classification tasks.

Nikki Mirghafori (Speech): speech processing, particularly speech and speaker recognition.


Sri Narayanan (AI): probabilistic models of language interpretation, graphical models of linguistic aspect, graphical models of stochastic grammars, semantics of linguistic aspect, on-line metaphor interpretation, and embodied rationality; more recently models of the role of sub-cortical structures (like basal ganglia-cortex loops) in attentional control.

Vern Paxson (Networking): intrusion detection; Internet measurement; measurement infrastructure; packet dynamics; self-similarity.

Barbara Peskin (Speech): speech processing, including recognition and tracking of speech and speakers.

Lokendra Shastri (AI): Artificial Intelligence, Cognitive Science, and Neural Computation: neurally motivated computational models of learning, knowledge representation and inference; rapid memory formation in the hippocampal system; inference with very large knowledge-bases; neural network models for speech recognition; inferential search and retrieval.

Scott Shenker (Networking): congestion control, internet topology, game theory and mechanism design, scalable content distribution architectures, and quality of service.

Elizabeth Shriberg (Speech): Modeling spontaneous conversation, disfluencies and repair, prosody modeling, dialog modeling, automatic speech recognition, utterance and topic segmentation, psycholinguistics, computational psycholinguistics. (Also with SRI International).

Andreas Stolcke (Speech): probabilistic methods for modeling and learning natural languages, in particular in connection with automatic speech recognition and understanding. (Also with SRI International).

Chuck Wooters (Speech): systems issues for speech processing, particularly for automatic speech recognition; universals in pronunciation modeling.
Part II

RESEARCH GROUP REPORTS

In 2003 we continued to progress in our traditional areas of strength, while beginning to explore newer areas. Our continued success with US Federal grants provided much of the support for this work. Among other things, the Networking group began having a greater emphasis on network security, and the speech group began new work in speaker segmentation and recognition. In this report, we have described our research in terms of the five principal research groups in place during 2003: Networking, Speech, AI, Algorithms, and BCIS. We will also describe other activities that did not fit neatly into these categories.

1 Research Group Highlights

The following are a selection of key achievements in our research groups for the year 2003, both in group development and in research per se. Although not a complete listing and, by necessity, quite varied given the differing approaches and topics of each group, it should nonetheless give the flavor of the efforts in the ICSI community for the last year.

1.1 Networking

- XORP: The eXtensible Open Routing Platform made four releases during the year and grew to support a wider range of features and platforms.

- Two new IETF standards in congestion control: HighSpeed TCP for Large Congestion Windows (Experimental) DCCP, a congestion-control method for UDP (standardization expected shortly).

- Worm research: DARPA report “Large Scale Malicious Code: A Research Agenda”; analysis of the Slammer worm, finding that it infected 75,000 hosts in under 10 minutes; “A Taxonomy of Computer Worms” published.

- Proposals for redesigning BGP, the interdomain routing protocol, that would provide scalability and robustness.

1.2 Algorithms

- Development of algorithms and software for key problems in functional genomics.

- Participation in creation of a UC Berkeley strategic initiative in computational biology, with seven new faculty positions.

- Addition of a research group in Portland, Oregon under Dr. John Moody, working in statistical learning theory and computational finance.
1.3 AI

- Selection of AI group for a larger role in Phase 2 of the ARDA Aquaint program.
- FrameNet funding from Darpa as a potential key resource in Intelligent Systems.
- Full day session on Embodied Construction Grammar at the International Conference on Cognitive Linguistics in Spain.
- First implementation of CPRM – coordinated probabilistic relation models.
- Narayanan paper on Semantic Web among the ten most cited.

1.4 Speech

- New awards from NSF and the EU; in particular, speaker recognition from NSF and Augmented Multiparty Interaction (AMI), a 6th framework Integrated Project, from the EU. Also, now part of a large UC Berkeley project to design computational aids for 3rd world users.
- The ICSI Meeting Corpus was released, both to our collaborators and (via an official LDC release) to the wider community of users. This is the largest carefully transcribed corpus of meetings currently available.
- Dr. Nikki Mirghafori joined the group as a research staff member.
- Successful completion of the English language component of SmartKom spoken language system.
- First successful application of long-time (500 ms) temporal features to conversational telephone speech recognition, as part of the DARPA EARS “Novel Approaches” project.

1.5 BCIS highlights

- Multi-year funding for BFOIT from the Bechtel Foundation.
- Anna Lee Saxenian appointed Dean of SIMS.
- Completion of several comparative studies, books in press.
2 Networking

2.1 Internet Architecture

2.1.1 Host-based Congestion Control

DCCP: Historically, the great majority of Internet unicast traffic has used congestion-controlled TCP, with UDP making up most of the remainder. UDP has mainly been used for short, request-response transfers, like DNS and SNMP, that wish to avoid TCP's three-way handshake, retransmission, and/or stateful connections. Recent years have seen the growth of applications that therefore use UDP in a different way. These applications, including RealAudio, Internet telephony, and on-line games such as Half Life, Quake, and Starcraft, share a preference for timeliness over reliability, and use UDP instead of TCP. This growth of long-lived non-congestion-controlled traffic, relative to congestion-controlled traffic, poses a real threat to the overall health of the Internet. To address this, we are continuing the design of a new protocol, Datagram Congestion Control Protocol (DCCP), that combines unreliable datagram delivery with built-in congestion control. DCCP is currently being standardized in the IETF. Continuing work on DCCP has included proposed modifications to the congestion control mechanisms to allow faster startup, faster recovery of old sending rates after idle periods, and fuller consideration of the problems introduced by variable packet sizes. To evaluate the protocol and investigate new, efficient user-kernel interaction mechanisms, we built a prototype DCCP API for Linux.

High Speed TCP: HighSpeed TCP is an experimental modification to TCP's congestion control mechanisms for better operation in TCP connections with large congestion windows. The congestion control mechanisms of the current Standard TCP constrains the congestion windows that can be achieved by TCP in realistic environments. For example, for a Standard TCP connection with 1500-byte packets and a 100 ms round-trip time, achieving a steady-state throughput of 10 Gbps would require an average congestion window of 83,333 segments, and a packet drop rate of at most one congestion event every 5,000,000,000 packets (or equivalently, at most one congestion event every 1 2/3 hours). This is widely acknowledged as an unrealistic constraint. To address this limitation of TCP, we proposed HighSpeed TCP, a modification to TCP's congestion control mechanisms which is now standardized as Experimental. HighSpeed TCP is under active investigation in the network research community, along with other proposals for high-performance TCP.

Measuring TCP's Loss Rate: The problem of measuring the loss rate experienced by a TCP connection is easy at first glance since TCP is a reliable transport that retransmits all lost data. However, in this work we find that simply using the number of retransmits represents only a very gross estimate of the loss rate due to TCP's use of a coarse-grained retransmission strategy that triggers spurious retransmissions in some situations. Estimating the total loss rate experienced by a TCP connection is useful for several reasons. First, passively observing TCP connections in networks is a powerful method to uncover network dynamics. Also, the loss rate has been shown to be a fundamental component of the ultimate performance attained by a TCP connection. Finally, techniques which attempt to untangle the root causes of loss (i.e., congestion-related loss or corruption-based loss) and change TCP's congestion response appropriately require an
 accurate measure of the total loss rate. Therefore, we introduce several algorithms that use hints in returning acknowledgments to determine which retransmits were sent needlessly. By making this determination we can refine the retransmission count into a more accurate loss estimate.

**Congestion Control in High BER Networks:** One of the lingering problems faced by transport protocols in networks with high corruption-based loss rates is that congestion control is invoked on each loss under the assumption that the loss was caused by contention for resources in routers. This assumption is conservative in its effort to avoid congestion collapse. Further, the assumption is generally correct in wireline networks. However, in wireless networks where packets are corrupted on a regular basis the assumption does not hold. The goal of this project is to untangle the cause of losses by using an accurate total loss estimate made by the transport sender with an estimate of the corruption-based loss rate provided by the intermediate nodes in the network (routers, wireless basestations, etc.). Using these two loss rates the transport sender is able to provide a congestion response that is both appropriate to the current level of congestion while not being hindered by corruption-based losses.

**Detecting Spurious Retransmits:** TCP and SCTP both provide reliability by retransmitting lost data. In addition, losses are taken as an indication of network congestion and used to trigger congestion control in the data sender (i.e., a reduction in the sending rate). By detecting spurious retransmits a TCP or SCTP sender can "undo" congestion control decisions that were unnecessary. We have been working on a scheme that uses TCP's DSACK option and SCTP's Duplicate TSN notifications to identify spurious retransmissions.

**Early Retransmit:** TCP retransmissions are triggered by either information in acknowledgments returned by the receiver or via a retransmission timeout (RTO). Using the RTO for loss recovery is costly from a performance standpoint. However, when the sending rate is low TCP relies on the RTO for most of its retransmissions. We introduce Early Retransmit to aid performance by triggering retransmits based on acknowledgments more aggressively when the sending rate is low than when the sending rate is high (and TCP is naturally more apt to use mechanisms other than the RTO for loss repair).

### 2.1.2 Router-based Congestion Control

**Selective Early Drop** The Selective Early drop (SED) pre-filter is an algorithm for selectively dropping packets from high-rate flows arriving at a router with a single congested link and a droptail output queue, eliminating both congestion and inequity in bandwidth distribution to flows. SED maintains a dynamically changing per-flow bandwidth limit, and maintains a count for each flow representing the extent to which the flow exceeds this limit. Simulation and formal analysis demonstrate that the number of positive counts will be small. Packets from a flow with a positive count are dropped with increasing severity as the count increases. We demonstrate through extensive ns simulations that, for a variety of combinations of TCP-compliant, constant-bit-rate and bursty flows, SED achieves high throughput and near-optimal fairness.
2.1.3 Routing

**Lightweight Security Mechanisms for BGP:** BGP, the current inter-domain routing protocol, assumes that the routing information propagated by routers is correct. A violation of this assumption leaves the current infrastructure vulnerable to misconfigurations and deliberate attacks that alter the behavior of the control and data planes. Deliberate attackers along a path can potentially render destinations unreachable, eavesdrop on data passing through them, impersonate an site, and take countermeasures against security measures. We developed a series of mechanisms of increasing complexity that deal with attacks of increasing sophistication. One mechanism involves probing of data paths. The other mechanisms involve comparing route information along multiple paths, using redundancy and cryptographic one-way functions instead of shared key cryptography to establish the validity of a route advertisement. Although these mechanisms do not achieve perfect security, they do provide much better security than what exists today. They are easily deployable and do not require a key distribution infrastructure. However, even these measures are not sufficient against colluding attackers; here, we must augment our arsenal with proposed changes to acceptable BGP policies.

**Incentive Mechanisms for BGP:** ASs currently use policy features in BGP to make sure that local routing decisions are consistent with their incentives. One could envision a more general accommodation of incentives, where ASs were recompensed for carrying traffic. This project explores how to create a strategy-proof form of BGP that does not require substantial additional overhead to compute payments.

**Traffic Engineering:** A major challenge for inter-domain traffic engineering is the level of uncertainty an ISP faces. An ISP has limited, and often inaccurate, knowledge about user demand, network topology, available network resources and routing policies in other domains that their traffic traverses. In this project we explicitly deal with such uncertainty in terms of search games, in which the network operator must balance loads across multiple paths on which the available capacity is unknown and time-varying. We develop algorithms for solving such search games based on binary congestion feedback from the network. The algorithms are provably near-optimal in the static case. Extensive simulations further demonstrate that they converge very quickly in a dynamic environment, are highly robust to measurement noise, and are quite stable even in the presence of multiple competing players.

**Towards a Next Generation Inter-domain Routing Protocol:** BGP, the current inter-domain routing protocol, is known to suffer from several pressing problems that render the current infrastructure vulnerable to attacks and threaten to impede Internet growth. This project proposes a new inter-domain routing protocol that combines features from both link-state and path-vector routing. This hybrid Link-state Path-vector protocol, called HLP, addresses five specific issues with BGP: lack of scalability, lack of security, poor convergence, lack of fault isolation, and lack of transparency for problem diagnosis. HLP preserves the basic operational and economic model of BGP and only modifies the way in which routing information is propagated. Our hope is that this proposal will stimulate a new debate about the nature of a next-generation BGP.
Selfish Routing in Internet-Like Environments: A recent trend in routing research is to avoid inefficiencies in network-level routing by allowing hosts to either choose routes themselves (e.g., source routing) or use overlay routing networks (e.g., Detour or RON). Such approaches result in selfish routing, because routing decisions are no longer based on system-wide criteria but are instead designed to optimize host-based or overlay-based metrics. A series of theoretical results showing that selfish routing can result in suboptimal system behavior have cast doubts on this approach. In this project, we use a game-theoretic approach to investigate the performance of selfish routing in Internet-like environments. We focus on intra-domain network environments and use realistic topologies and traffic demands in our simulations. We show that in contrast to theoretical worst cases, selfish routing achieves close to optimal average latency in such environments. However, such performance benefit comes at the expense of significantly increased congestion on certain links. Moreover, the adaptive nature of selfish overlays can significantly reduce the effectiveness of traffic engineering by making network traffic less predictable.

2.1.4 Novel Approaches

A Layered Naming Architecture for the Internet: Currently the Internet has only one level of name resolution, DNS, which converts user-level domain names into IP addresses. In this project we borrow liberally from the literature to argue that there should be three levels of name resolution: from user-level descriptors to service identifiers; from service identifiers to endpoint identifiers; and from endpoint identifiers to IP addresses. These additional levels of naming and resolution (1) allow services and data to be first class Internet objects and (2) facilitate mobility and provide an elegant way to integrate middleboxes into the Internet architecture. We further argue that flat names are a natural choice for the service and endpoint identifiers. Hence, this architecture requires scalable resolution of flat names, a capability that distributed hash tables (DHTs) can provide.

Untangling the Web from DNS: The Web relies on the Domain Name System (DNS) to resolve the hostname portion of URLs into IP addresses. This marriage-of-convenience enabled the Web's meteoric rise, but the resulting entanglement is now hindering both infrastructures: the Web is overly constrained by the limitations of DNS, and DNS is unduly burdened by the demands of the Web. There has been much commentary on this sad state-of-affairs, but dissolving the ill-fated union between DNS and the Web requires a new way to resolve Web references. To this end, this project describes the design and implementation of Semantic Free Referencing (SFR), a reference resolution infrastructure based on distributed hash tables (DHTs).

Routing as a Service: Many recent proposals have argued for giving end-hosts control over routing in the network to satisfy the growing demands of applications. However, these proposals either run at an overlay level independent of one another, or else lack support for end-hosts to discover the desired routes. In this project, we propose a network architecture that addresses these limitations. Our basic premise is that specialized route computation should be provided as a service and not embedded in the infrastructure. This design allows the routing functionality to evolve without changing the infrastructure. Our architecture consists of three entities: (a) a forwarding infrastructure that enables end-hosts to setup routes, (b) Routing Service (ROSE)
providers that compute routes (conforming to application requirements) based on network information that the infrastructure provides, and (c) end-hosts that setup the routes, obtained by querying ROSE, in the infrastructure. We address the issues of trust, scalability of the ROSE architecture, and deployability. We demonstrate the feasibility of our solution by conducting experiments on a prototype deployed on PlanetLab. To illustrate the benefits of our architecture we evaluate two applications: metric-sensitive multicast, and resilient routing.

**Host Mobility using i3:** We propose the Robust Overlay Architecture for Mobility (ROAM) to provide seamless mobility for Internet hosts. ROAM is built on top of the Internet Indirection Infrastructure (I3). With i3, instead of explicitly sending a packet to a destination, each packet is associated with an identifier. This identifier defines an indirection point in i3, and is used by the receiver to obtain the packet. ROAM takes advantage of end-host ability to control the placement of indirection points in i3 to provide efficient routing, fast handoff, and preserve location privacy for mobile hosts. In addition, ROAM allows end hosts to move simultaneously, and is as robust as the underlying IP network to node failure. We have developed a user-level prototype system on Linux that provides transparent mobility without modifying applications or the TCP/IP protocol stack. Simulation results show that ROAM's latency can be as low as 0.25-40% of Mobile IP. Experimental results show that with soft handoff the TCP throughput decreases only by 6% when there are as many as 0.25 handoffs per second.

### 2.2 Measurements and Modeling

**Models for network research:** By a network model for a simulation or experiment, we mean the full range of parameters that might affect a simulation or experiment: network topology, traffic generation, end-node protocol behavior, queue drop policies, congestion levels, and so forth. Network models used in practice often have little relationship to Internet reality, or an unknown relationship to Internet reality. We simply don't know whether the models we use are valid. This basic question has led to difficulties both in our own research and in our evaluation of other work. We are continuing a project to broaden discussion within the research community about the models we use, with the goal of a more agreed-upon set of modeling practices for specific research questions. Recent work in this area includes an evaluation of models for wireless links in terms of their effect on the performance of transport protocols, and tools to aid in characterizing congestion in the Internet.

**Measuring bottleneck capacities:** The best way to build and evaluate this kind of Internet model is through ongoing, large-scale Internet measurement. Application-centric measurement techniques are often unsuitable for this purpose: for example, long-lived, large-scale active measurements can generate too much traffic overhead. We are developing and/or adapting a comprehensive set of accurate, passive, trace-based measurement tools, focused on creating a more faithful description of the Internet. The first tools, MultiQ and Mystery, detect multiple bottleneck capacities and their relative order from a passive trace, and correlate this with information about loss events and RTT changes. These tools can evaluate historical changes in bottleneck capacity, and investigate other questions in depth, such as changes in the level of statistical multiplexing on bottlenecks.
NIMI: The NIMI (National Internet Measurement Infrastructure) project focuses on developing and deploying a system for facilitating coordinated measurement from a number of points around the Internet. Work is nearing completion on a major revision of the NIMI architecture which aims to refine its authorization, security, and resource control mechanisms.

Middlebox Performance: Middleboxes are intelligent entities that are appearing with increasing frequency as intermediate nodes in today’s Internet. These devices are designed to aid security (e.g., firewalls, traffic normalizers), performance (e.g., TCP acknowledgment spoofer), as well as extend the address space of the network (e.g., network address translators). While there are many legitimate reasons for adding a middlebox to the network there has been little investigation into the impact of middleboxes on performance and robustness in the wild. We study a particular middlebox infrastructure in place at a large laboratory and find performance costs and benefits. In addition, we find many new failure modes introduced by the middlebox, as well as increased instances of common failure modes. However, while failures increase in the presence of the middlebox they are still infrequent events (the failure rate with and without the middlebox is always less than 0.1% in our experiments).

2.3 Peer-to-Peer Systems

2.3.1 Distributed Hash Tables

IRIS Project: The Infrastructure for Resilient Internet Services (IRIS) project combines the efforts of 12 PIs from five institutions (ICSI, UCB, MIT, Rice, NYU). The IRIS project is developing a novel decentralized infrastructure, based on distributed hash tables (DHTs), that will enable a new generation of large-scale distributed applications. DHTs are robust in the face of failures, attacks and unexpectedly high loads. They are scalable, achieving large system sizes without incurring undue overhead. They are self-configuring, automatically incorporating new nodes without manual intervention or oversight. They provide a simple and flexible interface and are simultaneously usable by many applications.

Query Processing: The database research community prides itself on scalable technologies. Yet database systems traditionally do not excel on one important scalability dimension: the degree of distribution. This limitation has hampered the impact of database technologies on massively distributed systems like the Internet. To rectify this, we propose the initial design of PIER, a massively distributed query engine based on overlay networks, which is intended to bring database query processing facilities to new, widely distributed environments. We motivate the need for massively distributed queries, and argue for a relaxation of certain traditional database research goals in the pursuit of scalability and widespread adoption. We have simulation results showing PIER gracefully running relational queries across thousands of machines, and initial results from the same software base in actual deployment on a large experimental cluster.

Load Balancing: We propose an algorithm for load balancing in heterogeneous, dynamic peer-to-peer systems. The algorithm is based on a partition of the address space of distributed hash table (DHT) into regions assigned to virtual servers. Directories within the DHT maintain information about the loads (in bandwidth, storage or computation) imposed by the virtual
servers. Load balancing is achieved by adjusting the assignment of virtual servers to physical servers.

Extensive simulations show that, in the face of object arrivals, the algorithm moves only a negligible fraction of the arriving load. In a dynamic system where servers arrive and depart, the algorithm moves only a small fraction of the load that the underlying distributed hash table moves in processing server arrivals and departures.

Range Searches over DHTs: Distributed Hash Tables are scalable, robust, and self-organizing peer-to-peer systems that support exact match lookups. This project describes the design and implementation of a Prefix Hash Tree - a distributed data structure that enables more sophisticated queries over a DHT. The Prefix Hash Tree uses the lookup interface of a DHT to construct a trie-based structure that is both efficient (updates are doubly logarithmic in the size of the domain being indexed), and resilient (the failure of any given node in the Prefix Hash Tree does not affect the availability of data stored at other nodes).

Spurring the Adoption of DHTs through OpenHash: The past three years have seen intense research into Distributed Hash Tables (DHTs): both into algorithms for building them, and into applications built atop them. These applications have spanned a strikingly wide range, including file systems, event notification, content distribution, e-mail delivery, indirection services, web caches, and relational query processors. While this set of applications is impressively diverse, the vast majority of application building is done by a small community of DHT researchers. Why, then, has this community of developers remained narrow? First, keeping a research prototype of a DHT running continually requires effort, and experience with DHT code. Second, significant testbed resources are required to deploy and test DHT-based applications. Our central tenet is that we, as a community, need to harness the ingenuity and talents of the vast majority of application developers who reside outside the rarified but perhaps sterile air of the DHT research community. To that end, we issue a call-to-arms to deploy an open, publicly accessible DHT service that would allow new developers to experiment with DHT-based applications without the burden of deploying and maintaining a DHT. We call this system OpenHash.

The Impact of DHT Routing Geometry on Resilience and Proximity: The various proposed DHT routing algorithms embody several different underlying routing geometries. These geometries include hypercubes, rings, tree-like structures, and butterfly networks. In this project we focus on how these basic geometric approaches affect the resilience and proximity properties of DHTs. One factor that distinguishes these geometries is the degree of flexibility they provide in the selection of neighbors and routes. Flexibility is an important factor in achieving good static resilience and effective proximity neighbor and route selection. Our basic finding is that, despite our initial preference for more complex geometries, the ring geometry allows the greatest flexibility, and hence achieves the best resilience and proximity performance.

2.3.2 Other P2P Systems

Making Gnutella Scalable: DHTs represent a somewhat radical departure from the more ad-hoc nature of deployed systems such as Gnutella. The goal of this project is to explore P2P design alternatives that achieve scalability while retaining the ad-hoc nature of systems such as
Gnutella and KaZaa. This approach represents an incremental change to the deployed infrastructure rather than a wholesale replacement as required by DHTs and thus might face an easier path to adoption. A design that incorporates flow control, one-hop replication, and capacity-sensitive topology adaptation achieves performance that is roughly four orders of magnitude better than Gnutella-like designs.

2.4 Security, Malware, and Intrusion Detection

Analyzing the threat of Internet worms: The ability of attackers to rapidly gain control of vast numbers of Internet hosts using automated “worms” poses an immense risk to the overall security of the Internet. This project analyzes the threat posed by current and future worms, and possible counter-measures. Our measurements and analysis with colleagues of the January 2003 “Slammer” worm revealed that it was by far the fastest-spreading worm seen to date, infecting 75,000 Internet hosts in under 10 minutes. We have completed writing an extensive research agenda for “large scale malicious code,” formulated a defensible estimate that a plausible “worst case” worm might be capable of inflicting $100 billion in economic damage to the United states, and developed a taxonomy for characterizing different types of worms in terms of how they spread, their attack “payloads,” possibilities of detection, and classes and motivations of potential worm authors.

Semantic packet trace transformation: No Internet site is ever willing to release traces of its traffic if those traces include packet contents, due to security and privacy concerns. Yet without such traces, it is exceedingly difficult to accurately evaluate the effectiveness of various intrusion detection algorithms. This project developed a framework by which the packet contents of traces can be altered in a semantic context: for example, by correctly recognizing when particular bytes in a packet payload correspond to a user name, file name, password, email header or body, web request or response, etc. The goal is to encourage the public release of actual, large-scale traces. The trace transformation also has other applications, such as greatly reducing the size of archived traces by trimming the contents of large connections while keeping small connections intact. We developed anonymizers for the FTP, HTTP, SMTP, IDENT, andfinger protocols. The first of these comprised the main effort, which culminated in a release of packet traces—including contents—of ten days’ worth of anonymous FTP traffic recorded at the Lawrence Berkeley National Laboratory. We believe this is the first-ever public release of a significant set of traces that include packet contents derived from regular user traffic.

Methodologies for assessing Internet-scale security analyses and defenses: a major problem with Internet-scale security research is understanding how to soundly evaluate the accuracy of the research. For example, how can we corroborate studies that purport to model the behavior of global-scale worms? Or, given a proposed defense mechanism for protecting against distributed denial-of-service (DDOS) attacks, how can we judge its likely efficacy when deployed in the operational Internet? This new collaborative effort aims to advance the state of the art in such assessment of Internet-scale security research. Much of the work is in conjunction with the DETER project, which is building a large testbed for conducting large-scale Internet security experiments in a controlled setting. Our current efforts focus on exploring notions of “scale down”: what sort of abstractions can be soundly employed to model the Internet on a tractably reduced scale?
Contextual signatures for intrusion detection: Many network intrusion detection systems, including the most popular commercial and freeware ones, are oriented around matching “signatures” in packet contents or connection byte streams: looking for exact matches of specific strings. Signature-matching often suffers from a high false positive rate due to the absence of being able to incorporate additional context into the matching decisions. We have developed a signature-matching engine with the key property that it is integrated into the Bro intrusion detection system’s “event engine” framework. Doing so allows us to augment the low-level nature of signature-matching with Bro’s powerful contextual analysis capabilities. This work is now ready for operational deployment at the Lawrence Berkeley National Laboratory and at the Technical University of Munich.

Architecting “independent state” for network intrusion detection systems: Network intrusion detection systems (NIDS) rely on managing a great amount of state for tracking active connections and the specifics of behavior observed in the past. Often much of this state resides solely in the volatile processor memory accessible to a single user-level process on a single machine. In this work we have developed with colleagues an architecture that facilitates independent state, i.e. internal fine-grained state that can be propagated from one instance of a NIDS to others running either concurrently or subsequently. Our unified architecture offers a wealth of possible applications that hold the potential to greatly enhance the power of a NIDS; we are exploring examples such as distributed processing, load parallelization, sharing attack information between different sites, controlling loss of state across restarts, dynamic reconfiguration, high-level policy maintenance, and support for profiling and debugging.

Research coupled with operational intrusion detection: There is a world of difference between intrusion detection research as explored in a computer science department lab versus the real-world problems encountered with 24x7 intrusion detection operation at a busy site. This ongoing project, in collaboration with the System and Network Security groups at the Lawrence Berkeley National Laboratory and at the University of California, Berkeley, centers on research and development in support of the 24x7 use of the Bro intrusion detection system as a primary component of site security at those institutions.

Detecting portscanners: Attackers routinely perform random “portscans” of Internet addresses to find vulnerable servers to compromise. Network intrusion detection systems (NIDS) attempt to detect such behavior and flag these portscanners as malicious. An important need in such systems is prompt response: the sooner a NIDS detects malice, the lower the resulting damage. At the same time, a NIDS should not falsely implicate benign remote hosts as malicious. Balancing the goals of promptness and accuracy in detecting malicious scanners is a delicate and difficult task. With colleagues we have developed a connection between this problem and the theory of sequential hypothesis testing and showed that one can model accesses to local Internet addresses as a random walk on one of two stochastic processes, corresponding respectively to the access patterns of benign remote hosts and malicious ones. The detection problem then becomes one of observing a particular trajectory and inferring from it the most likely classification for the remote host. We used this insight to develop Threshold Random Walk (TRW), a novel on-line detection algorithm that identifies malicious remote hosts. Using an analysis of traces from two qualitatively different sites, we showed that TRW requires
a much smaller number of connection attempts (4 or 5 in practice) to detect malicious activity compared to previous schemes, while also providing theoretical bounds on the low (and configurable) probabilities of missed detection and false alarms.

**Detecting relays:** One common form of abuse to which attackers subject Internet hosts is manipulating the host to serve as a *relay* that echoes traffic incoming to the host back out to other hosts. Relays can manifest in a number of ways, such as: open Web proxies, email servers exploited for propagating “spam,” “stepping stones” with which attackers can launder their tracks to thwart discovery of their identity, “reflectors” that serve to make DDOS attacks much more diffuse, and Internet Relay Chat servers which attackers establish on compromised hosts both as a means to brag to their community about having broken into the host, and, increasingly, as a control system for “botnet” networks of large numbers of “zombie hosts” that can be commanded on demand to launch massive flooding attacks. In addition, Internet *worms* constitute another form of relay, where the relay is comprised directly of the propagation of the worm, first into the local host and then back out to other hosts as the worm spreads further. In this new collaborative project we are aiming to devise lightweight, generic algorithms for detecting hosts being abused as relays by observing similarities between content sent to a host and content which the host then transmits a short time later. We have developed a preliminary proof-of-principle algorithm but significant hurdles remain in refining it to achieve acceptably high performance, and developing an accompanying policy component to deal with the large number of benign relays that occur in large traffic streams in practice.

**Integrating traffic sampling into intrusion detection:** Techniques to sample network traffic have seen a flurry of recent advancement in support of network measurement. On the other hand, historically sampling has been viewed as of little utility for network intrusion detection because attacks are generally a minor component of a traffic stream, and thus sampling that traffic stream is likely to diminish the available analysis signal rather than augment it. We have begun investigating the application of sampling techniques to enhance intrusion detection. First, we are looking at ways to characterize different forms of “large” traffic flows. Some of these flows are of direct interest for detecting attacks—for example, rapidly discovering traffic floods can enable operators to take steps to ameliorate both damage to the victim and also excessive load on the intrusion detection system’s analysis components. In addition, a general strategy we pursue in our network intrusion detection research is to find efficient mechanisms for detecting activity expressed in more abstract terms, whether benign or malicious. Such information can often augment the power of high-level security analysis by providing additional context. More generally, while much traffic sampling focuses on *randomized* techniques, we are investigating enhancing the standard “packet filter” mechanism operating systems provide by supplementing the filter with more powerful basic operators, including access to significant additional *state* in the form of associative tables. In doing so we aim to support both statistical sampling such as for flood-detection, but also fine-grained, detailed analysis of specific, individual traffic sources.

**Worm Detection and Suppression:** Worms, self propagating network programs, represent a substantial threat to our computing infrastructure. By first developing a taxonomy of worms, we are then able to construct defenses against particular classes of worms. Our first proposals are substantial improvements to scanning-worm containment, including an algorithm capable
of quickly detecting scanning while being suitable for efficient hardware implementations, the use of communication to further increase sensitivity, and possible attacks (and counterattacks) for when worm authors wish to avoid containment systems.

**EMIST Worm Modeling:** The development of the Deter testbed requires models of malicious behavior in order to evaluate defenses. We are focusing in particular on the behavior of Slammer, which was able to spread across the Internet in less than 10 minutes. By being able to model Slammer in the Deter testbed, we will be able to then test defenses against this particularly virulent class of worm.

### 2.5 Extensible Open Router Platform

**XORP:** Network researchers face a significant problem when deploying software in routers, either for experimentation or for pilot deployment. Router platforms are generally not open systems, in either the open-source or the open-API sense. The eXtensible Open Router Platform (XORP) attempts to address these issues. Key goals are extensibility, performance and robustness. We show that different parts of a router need to prioritize these differently, and examine techniques by which we can satisfy these often conflicting goals. We aim for XORP to be both a research tool and a stable deployment platform, thus easing the transition of new ideas from the lab to the real world. XORP made four releases during the year and grew to support a wider range of features and platforms.

**Click Modular Router:** This project is creating an extensive modular toolkit for building routers and other packet processors. Click will act as XORP's forwarding path. Work continues on IPv6 and multicast support. Click is also well suited for trace analysis and measurement; we are building an extensive set of elements that make measurements both easy and fast.

### 2.6 Sensornets

**Geographic Routing without Location Information:** For many years, scalable routing for wireless communication systems was a compelling but elusive goal. Recently, several routing algorithms that exploit geographic information (e.g., GSPR) have been proposed to achieve this goal. These algorithms refer to nodes by their location, not address, and use those coordinates to route greedily, when possible, towards the destination. However, there are many situations where location information is not available at the nodes, and so geographic methods cannot be used. In this project we define a scalable coordinate-based routing algorithm that does not rely on location information, and thus can be used in a wide variety of ad hoc and sensornet environments.

**Data-Centric Storage in SensorNets:** Previous work such as directed diffusion has identified data-centric routing as a scalable dissemination mechanism for large-scale sensornets. This project explores the utility of a companion method, data-centric storage. Data-centric storage provides a way to scalably support a broad range of sensornet queries.
**Geographic Hash Tables:** In this project we developed GHT, a Geographic Hash Table system, that implements DCS on sensornets. GHT hashes keys into geographic coordinates, and stores a key-value pair at the sensor node geographically nearest the hash of its key.

**A Distributed Index for Features in Sensor Networks (DIFS):** Sensor networks pose new challenges in the collection and distribution of data. Recently, much attention has been focused on standing queries that use in-network aggregation of time series data to return data statistics in a communication-efficient manner. In this work, rather than consider searches over time series data, we consider searches over semantically rich high-level events, and present the design, analysis, and numerical simulations of a spatially distributed index that provides for efficient index construction and range searches. The scheme provides load balanced communication over index nodes by using the governing property that the wider the spatial extent known to an index node, the more constrained is the value range covered by that node.

**Time Synchronization:** The recently proposed Reference-Broadcast Synchronization proposal provides on-demand pairwise synchronization of sensornet nodes with low overhead and high precision. In this project we introduce a model of the basic RBS synchronization paradigm, characterize the optimally precise clock synchronization algorithm, and establish its global consistency. We also discuss synchronization in the presence of clock skews. In the course of this analysis we point out unexpected connections between optimal clock synchronization, random walks and resistive networks.

### 2.7 Internet Community Activities

ICSI researchers are quite active in the Internet research community. In addition to the normal academic duties of serving on program committees and editorial boards, ICSI researchers devote substantial time to more practical duties associated with the Internet Engineering Task Force (IETF) and Internet Research Task Force (IRTF). In particular, Vern Paxson is currently the Chair of the IRTF. Sally Floyd is currently a member of the Internet Architecture Board (IAB), which is a technical advisory board to the Internet Society and the IETF. Mark Allman chairs the IRTF's Internet Measurement Research Group (IMRG) and also participates in an effort to repair IETF's internal process.

### References

### Papers


Internet Drafts


RFCs


In Submission


3 Algorithms

3.1 Overview

The year 2003 was marked by an acceleration of the Algorithms Group's activities in computational biology and by a new venture in machine learning and computational finance led by Dr. John Moody.

The principal members of the ICSI Algorithms Group at Berkeley during 2003 were Group Leader Richard Karp, Professor Elwyn Berlekamp, postdoctoral fellows Jens Gramm, Eran Halperin, Robert Karanuthgam, Till Nierhoff, Roded Sharan and Till Tantau, and Ph.D. student Eric Xing. The affiliated activity in Portland, Oregon was led by Dr. John Moody and included postdoctoral fellow Yufeng Liu, Ph.D. student Matthew Saffel and research programmer Aron Rempel.

In computational biology we developed algorithms and publicly available software for a variety of key problems in functional genomics and systems biology, including discovery of conserved protein interaction networks and protein complexes, discovery of transcription factor binding sites and cis-regulatory modules, single-nucleotide polymorphism genotyping and haplotype phasing. Prof. Karp participated in a team effort that led to the designation of computational biology as a strategic initiative of the Berkeley campus, with seven new faculty positions to be allocated over the next five years.

Dr. Moody leads a group working on machine learning, neural and statistical computing, time series analysis, and computational finance. This activity is currently based in Portland Oregon.

Prof. Berlekamp continued his work on combinatorial game theory.

The Algorithms Group continues to work in its traditional areas of combinatorial algorithms and computational complexity, and to contribute to algorithmic developments in the ICIR Networking Group.

3.2 Computational Biology

3.2.1 Systems Biology

The dissection of complex biological systems is a challenging task, made difficult by the size of the underlying molecular network and by the heterogeneous nature of the control mechanisms involved. Novel high throughput techniques are generating massive datasets on various aspects of such systems. We have performed the first analysis of a highly diverse collection of genome-wide datasets, including gene expression, protein interactions, growth phenotype data and transcription factor binding, in order to reveal the modular organization of the yeast system [60]. By integrating experimental data of heterogeneous sources and types, we are able to perform analysis on a much broader scope than previous studies. At the core of our methodology is the ability to identify modules, namely, groups of genes with statistically significant correlated behavior across diverse data sources. Numerous biological processes are revealed through these modules, which also obey global hierarchical organization. We use the identified modules to study the yeast transcriptional network and also to predict the function of over 800 uncharacterized genes. Our analysis framework, entitled SAMBA, enables the processing of current as well as future sources of biological information, and is readily extendable to novel experimental techniques and higher organisms.
Deciphering the mechanisms that control gene expression in the cell is a fundamental question in molecular biology. This task is complicated by the large number of possible regulation relations in the cell, and the relatively small amount of available experimental data. Recently, a new class of regulation functions called chain functions was suggested. Many signal transduction pathways can be accurately modeled by chain functions, and the restriction to chain functions greatly reduces the vast search space of regulation relations. We have studied the computational problem of reconstructing a chain function using a minimum number of experiments, in each of which only few genes are perturbed. We give optimal reconstruction schemes for several scenarios and show their application in reconstructing the regulation of galactose utilization in yeast [19].

We have implemented a strategy for aligning two protein-protein interaction networks that combines interaction topology and protein sequence similarity to identify conserved interaction pathways and complexes [40]. Using this approach we show that the protein-protein interaction networks of two distantly related species, *Saccharomyces cerevisiae* and *Helicobacter pylori*, harbor a large complement of evolutionarily conserved pathways, and that a large number of pathways appears to have duplicated and specialized within yeast. Analysis of these findings reveals many well characterized interaction pathways as well as many unanticipated pathways, the significance of which is reinforced by their presence in the networks of both species.

Mounting evidence shows that many protein complexes are conserved in evolution. We have used conservation to find complexes that are common to yeast *S. cerevisiae* and bacteria *H. pylori* [36]. Our analysis combines protein interaction data, that are available for each of the two species, and orthology information based on protein sequence comparison. We develop a detailed probabilistic model for protein complexes in a single species, and a model for the conservation of complexes between two species. Using these models, one can recast the question of finding conserved complexes as a problem of searching for heavy subgraphs in an edge- and node-weighted graph, whose nodes are orthologous protein pairs. We tested this approach on the data currently available for yeast and bacteria and detected 11 significantly conserved complexes. Several of these complexes match very well with prior experimental knowledge on complexes in yeast only, and serve for validation of our methodology. The complexes suggest new functions for a variety of uncharacterized proteins. By identifying a conserved complex whose yeast proteins function predominantly in the nuclear pore complex, we propose that the corresponding bacterial proteins function as a coherent cellular membrane transport system. We also compare our results to two alternative methods for detecting complexes, and demonstrate that our methodology obtains a much higher specificity.

Our strategies for aligning protein-interaction networks and finding conserved protein complexes are implemented in PathBLAST, a network alignment and search tool for comparing protein interaction networks across species to identify protein pathways or complexes that have been conserved by evolution [41]. The basic method searches for high-scoring alignments between pairs of protein interaction paths, for which proteins of the first path are paired with putative orthologs occurring in the same order in the second path. This technique discriminates between true- and false-positive interactions and allows for functional annotation of protein interaction pathways based on similarity to the network of another, well-characterized species. PathBLAST is available at http://www.pathblast.org/ as a web-based query. In this implementation, the user specifies a short protein interaction path for query against a target protein-protein interaction network selected from a database. PathBLAST returns a ranked list of matching paths from the target network along with a graphical view of these paths and the
overlap among them. Target protein-protein interaction networks are currently available for H. pylori, S. cerevisiae, C. elegans, and D. melanogaster. Just as BLAST enables rapid comparison of protein sequences between genomes, tools such as PathBLAST are enabling comparative genomics at the network level.

3.2.2 Genome Rearrangements

One of the most promising ways to determine evolutionary distance between two organisms is to compare the order of appearance of orthologous genes in their genomes. The resulting genome rearrangement problem calls for finding a shortest sequence of rearrangement operations that sorts one genome into the other. We have provided a 1.5-approximation algorithm for the problem of sorting by transpositions and transreversals, improving on the 1.75 known ratio for this problem [35].

3.2.3 SNP Genotyping:

We have studied a design and optimization problem that occurs, for example, when single nucleotide polymorphisms (SNPs) are to be genotyped using a universal DNA tag array [8, 7]. The problem of optimizing the universal array to avoid disruptive cross-hybridization between universal components of the system was addressed in previous work. Cross-hybridization can, however, also occur assay-specifically, due to unwanted complementarity involving assay-specific components. Here we examine the problem of identifying the most economic experimental configuration of the assay-specific components that avoids cross-hybridization. Our formalization translates this problem into the problem of covering the vertices of one side of a bipartite graph by a minimum number of balanced subgraphs of maximum degree 1. We show that the general problem is NP-complete. However, in the real biological setting the vertices that need to be covered have degrees bounded by $d$. We exploit this restriction and develop an $O(d)$-approximation algorithm for the problem. We also give an $O(d)$-approximation for a variant of the problem in which the covering subgraphs are required to be vertex-disjoint. In addition, we propose a stochastic model for the input data and use it to prove a lower bound on the cover size. We complement our theoretical analysis by implementing two heuristic approaches and testing their performance on synthetic data as well as on simulated SNP data [7].

A generic genotyping assay utilizes a fixed set of reagents, which is independent of the actual target sample, to determine all present alleles. An example is the interrogation of several amplicons spanning polymorphic sites using an all $k$-mer array. Due to the high cost associated with a genotyping experiment, it is desirable to design a set of experiments, which maximizes the number of SNPs that can be genotyped in parallel per assay. We have investigated algorithmic approaches for optimally multiplexing SNP genotyping using generic assays [55]. We devise a graph theoretic formulation of the problem and use it to derive an approximation algorithm for the problem, and several practical heuristics. We apply our methods to simulated and real data, for evaluating the multiplexing rates afforded by generic techniques. The results on real human data show the practicality of generic approaches for genotyping, allowing, e.g., the genotyping of 5000 SNPs using four all 7-mer arrays.

The completion of the human genome makes it possible to look for telltale differences between the DNA sequences of different individuals and associate genomic variation with medical
condition. Key to association studies is the development of efficient and affordable genotyping techniques. Generic genotyping assays are independent of the target SNPs and offer great flexibility in the genotyping process. Efficient use of such assays calls for identifying sets of SNPs that can be interrogated in parallel under constraints imposed by the genotyping technology. We have studied problems arising in the design of genotyping experiments using generic assays. Our problem formulation deals with two main factors that affect the genotyping cost: The number of assays used and the number of PCR reactions required. We prove that the resulting computational problems are hard, but provide approximate and heuristic solutions to these problems, based on a graph representation of the input data. We tested our algorithmic approaches on an extensive collection of synthetic data and on data that was simulated using real SNP sequences. Our results show that the algorithms achieve optimal or near-optimal designs in many cases, and demonstrate the applicability of generic assays to SNP genotyping [23]

3.2.4 Gene Expression Analysis

We have developed a new Java-based graphical tool, called EXPANDER (EXPression ANAlyizer and DisplayER), for gene expression analysis and visualization [57]. This software contains several clustering methods including CLICK, K-Means, hierarchical clustering and self organizing maps, all controlled via a graphical user interface. It enables visualizing the raw expression data and the clustered data in several ways, as well as single-cluster and all-clusters evaluations via fitness scores and functional enrichment tests.

A central step in the analysis of gene expression data is the identification of groups of genes that exhibit similar expression patterns. Typically, different clustering algorithms yield different clustering solutions on the same data, and there is no agreed upon guideline for choosing among them. We have developed a novel statistically-based method for assessing a clustering solution according to prior biological knowledge. Our method can be applied to compare between different clustering solutions or to optimize the parameters of a clustering algorithm. The method is based on projecting vectors of biological attributes of the clustered elements onto the real line, such that the ratio of between-groups variance estimator and within group variance estimator is maximized. The projected data is then scored using a non-parametric analysis of variance test, and the score's sensitivity is evaluated. We validated our approach using simulated data and showed the superiority of our scoring method over extant ones, including separation to homogeneity ratio and the silhouette measure. Finally, we applied our method to evaluate popular clustering methods on a yeast cell-cycle gene expression dataset [20].

3.2.5 Transcriptional Regulation

Dissection of regulatory networks that control gene transcription is one of the greatest challenges of functional genomics. Transcription regulation is mediated primarily by combinatorial interactions between protein regulators called transcription factors (TFs) and their corresponding cis-regulatory recognition sites on the non-coding genomic sequences, often referred to as DNA motifs. We have addressed the problem of modeling generic features of structurally but not textually related DNA motifs, that is, motifs whose consensus sequences are entirely different, but nevertheless share “meta-sequence features” reflecting similarities in the DNA binding domains of their associated protein recognizers. We have developed LOGOS, a modular
Bayesian model for de novo motif detection [66] and Motif Prototyper [64], a profile Bayesian model which can capture structural properties typical of particular families of motifs.

By utilizing human genomic sequences, models for binding sites of known transcription factors and gene expression data, we have demonstrated that the reverse engineering approach, which infers regulatory mechanisms from gene expression patterns, can reveal transcriptional networks in human cells. To date, such methodologies have been successfully demonstrated only in prokaryotes and low eukaryotes. We developed computational methods for identifying putative binding sites of transcription factors and for evaluating the statistical significance of their prevalence in a given set of promoters. Focusing on transcriptional mechanisms that control cell cycle progression, our computational analyses revealed eight transcription factors whose binding sites are significantly over-represented in promoters of genes whose expression is cell cycle dependent. The enrichment of some of these factors is specific to certain phases of the cell cycle. In addition, we examined whether a higher order of organization could be revealed for these over-represented transcription factors. Several pairs of these transcription factors show a significant co-occurrence rate in cell cycle-regulated promoters. Each such pair suggests functional cooperation between its members in regulating the transcriptional program associated with cell cycle progression [13].

Gene transcription is controlled primarily via regulatory sequence elements that are recognized and bound by transcription factors. Recent findings suggest a modular organization of binding sites for transcription factors that cooperate in the regulation of genes. We have established a framework for finding recurrent cis-regulatory modules in the promoters of a selected set of genes and scoring their statistical significance. Proceeding from a database of identified binding site motifs and their genomic locations we seek motifs whose frequency in the selected promoters is different than in a background promoter set. We present several statistical tests designed for this purpose. We provide a hashing algorithm for detecting combinations of these motifs that co-occur in clusters within the selected promoters. The significance of such co-occurrences is evaluated using novel statistical scores. Our methods are combined in CREME, a suite of software which includes a browser for viewing the pattern of occurrence of selected cis-regulatory modules [54]. CREME relies on a database of putative transcription factor binding sites that have been annotated across the human genome using a library of position weight matrices and evolutionary conservation with the mouse and rat genomes. A search algorithm is applied to this dataset to identify combinations of transcription factors whose binding sites tend to co-occur in close proximity predominantly in the promoter regions of the input gene set. The identified cis-regulatory modules are statistically scored and significant combinations are reported and graphically visualized. Our web-server is available at http://creme.dcode.org/.

We applied our methodology to find modules within human-mouse conserved promoter segments, focusing on cell cycle regulated genes and stress response related genes. To validate the biological significance of the identified modules we tested whether the associated genes tended to be co-expressed or share similar function. In the cell cycle set roughly half the identified sets of genes were coherently expressed. On the stress response data about half of the detected sets fell predominantly into well-defined functional sub-categories [58].

We have proposed a method for identifying modules of transcription factor binding sites in a set of co-regulated genes, using only the raw sequence data as input. Our method is based on a novel probabilistic model that describes the mechanism of cis-regulation, including the binding sites of cooperating transcription factors, the organization of these binding sites into
short sequence modules, and the regulation of a gene by its modules. We show that our method is successful in discovering planted modules in simulated data and known modules in yeast. More importantly, we applied our method to a large collection of human gene sets, and found 83 significant cis-regulatory modules, which included 36 known motifs and many novel ones. Thus, our results provide one of the first comprehensive compendiums of putative cis-regulatory modules in the human genome [13].

3.2.6 Haplotyping

The study of haplotypes and their diversity in a population is central to disease association research. The problem of haplotyping via perfect phylogeny has received a lot of attention lately due to its applicability to real haplotyping problems and its theoretical elegance. In [15, 16] we developed an efficient algorithm for determining whether a set of genotypes could have been generated by a perfect phylogeny. This work is extended in [32] to the case where most, but not all, of the given genotypes come from a perfect phylogeny. The resulting algorithms have been implemented and are available at the HAP webserver (http://www.cs.columbia.edu/compbio/hap/).

The work on genotyping using a perfect or near-perfect phylogeny left two main issues open: the complexity of haplotyping with missing data, and whether the problem is linear-time solvable. In [28] we settle the first question and make progress toward answering the second one. Specifically, we prove that Perfect Phylogeny Haplotyping with missing data is NP-complete even when the phylogeny is a path and only one allele of every polymorphic site is present in the population in its homozygous state. Our result implies the hardness of several variants of the missing data problem, including the general Perfect Phylogeny Haplotyping Problem with missing data, and Hypergraph Tree Realization with missing data. On the positive side, we give a linear-time algorithm for Perfect Phylogeny Haplotyping when the phylogeny is a path. This variant is motivated by the abundance of yin-yang haplotypes in the genome. Our algorithm relies on a reduction of the problem to that of deciding whether a partially ordered set has width 2.

We have studied several problems arising in haplotype block partitioning. Our objective function is the total number of distinct haplotypes in blocks. We show that the problem is NP-hard when there are errors or missing data, and provide approximation algorithms for several of its variants. We also give an algorithm that solves the problem with high probability under a probabilistic model that allows noise and missing data. In addition, we study the multi-population case, where one has to partition the haplotypes into populations and seek a different block partition in each one. We provide a heuristic for that problem and use it to analyze simulated and real data. On simulated data, our blocks resemble the true partition more than the blocks generated by the LD-based algorithm of Gabriel et al. On single-population real data, we generate a more concise block description than extant approaches, with better average LD within blocks. The algorithm also gives promising results on real 2-population genotype data [43].

We have proposed a novel approach for phasing genotypes over long regions, which is based on combining information from local predictions on short, overlapping regions [17]. The phasing is done in a way which maximizes a natural maximum likelihood criterion. Among other things, this criterion takes into account the physical length between neighboring SNPs. We further give a confidence score to each position of the prediction, and use correlation information from the entire population to correct low confidence predictions. We evaluated our algorithm
on two real datasets using two different measures. Our results demonstrate the effectiveness of the approach. In all our tests we significantly outperformed the PHASE method. Our method is publicly available via a webserver at http://www.calit2.net/compbio/hap/.

The problem of inferring haplotypes from genotypes of single nucleotide polymorphisms (SNPs) is essential for the understanding of genetic variation within and among populations, with important applications to the genetic analysis of disease propensities and other complex traits. The problem can be formulated as a mixture model, where the mixture components correspond to the pool of haplotypes in the population. The size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. Thus methods for fitting the genotype mixture must crucially address the problem of estimating a mixture with an unknown number of mixture components. We have presented a Bayesian approach to this problem based on a nonparametric prior known as the Dirichlet process [65]. The model also incorporates a likelihood that captures statistical errors in the haplotype/genotype relationship, trading off these errors against the size of the pool of haplotypes. We describe an algorithm based on Markov chain Monte Carlo for posterior inference in our model. The overall result is a flexible Bayesian method that is reminiscent of parsimony methods in its preference for small haplotype pools. We apply our approach to the analysis of both simulated and real genotype data, and compare to extant methods.

3.2.7 Phylogeny

There are a few instances in which positive Darwinian selection has been convincingly demonstrated at the molecular level. We have presented a novel test for detecting excess of radical replacements at the amino-acid level [49]. Such excess can indicate the presence of positive Darwinian selection forces, but may also be the result of relaxed functional constraints or model misspecification. In this test, amino-acid replacements are characterized in terms of chemical distances, i.e., degrees of dissimilarity between the exchanged residues in a protein. The test identifies statistically significant deviations of the mean chemical distance from the random expectation, either along a phylogenetic lineage or across a subtree. The mean inferred distance is calculated as the average chemical distance over all possible ancestral sequence reconstructions weighted by their likelihood. Our method substantially improves over previous approaches by taking into account the stochastic process, tree phylogeny, among-site rate variation, and alternative ancestral reconstructions. We provide a fast linear time algorithm for applying this test to all branches and all subtrees of a given phylogenetic tree. We validate this approach by applying it to two well-studied datasets, the MHC class I glycoproteins serving as a positive control, and the house-keeping gene carbonic anhydrase I serving as a negative control.

Perfect phylogeny is one of the fundamental models for studying evolution. In [48] we have investigated the following variant of the model: The input is a species-characters matrix. The characters are binary and directed, i.e., a species can only gain characters. The difference from standard perfect phylogeny is that for some species the states of some characters are unknown. The question is whether one can complete the missing states in a way that admits a perfect phylogeny. The problem arises in classical phylogenetic studies, when some states are missing or undetermined. Quite recently, studies that infer phylogenies using inserted repeat elements in DNA gave rise to the same problem. Extant solutions for it take time $O(n^2m)$ for $n$ species and $m$ characters. We provide a graph theoretic formulation of the problem as a graph sandwich problem, and give near-optimal $O(nm)$-time algorithms for the problem. We also study the
problem of finding a single, general solution tree, from which any other solution can be obtained by node splitting. We provide an algorithm to construct such a tree, or determine that none exists.

In [29] we consider the reconstruction of perfect phylogenies from binary character data with missing values, and related problems of inferring complete haplotypes from haplotypes or genotypes in which some data is missing. In cases where the problems considered are \textit{NP}-hard we assume a plausible \textit{rich data hypothesis} under which they become tractable.

### 3.3 Risk, Reward, Reinforcement and Computational Finance

Professor John Moody joined ICSI on September 2, 2003. His research interests include machine learning, neural and statistical computing, time series analysis, and computational finance. He held previous faculty appointments in computer science at Oregon Graduate Institute (OGI) and Yale University, did his postdoc at the Institute for Theoretical Physics at UCSB and earned his Ph.D. in theoretical physics at Princeton University.

For the 2003-2005 academic years, John is working part-time at ICSI and establishing an off-site laboratory in the “silicon forest” area to the west of Portland, Oregon. The lab’s neighbors include major campuses of Intel, Tektronix, HP, Mentor Graphics, Xerox, NEC, Applied Materials, Sharp and OGI. He is taking a long-awaited break from full-time academia to devote most of his time to his entrepreneurial efforts at J E Moody & Company LLC, a quantitative investment management firm that uses machine learning methods to trade the currency, fixed-income and commodity markets.

Moody is currently directing two projects at ICSI:

- “ITR: Risk, Reward and Reinforcement”, NSF IIS-0342634.
- “Research Laboratory for Computational Finance”, funded by a donation from the Gordon / Petty Isis Foundation with matching funds from Intel.

Under NSF sponsorship, he hired Postdoctoral Fellow Yufeng Liu, a recent Biophysics Ph.D. from Carnegie Mellon University who also earned his M.S. in Machine Learning with research in robotics at CMU. Yufeng is working on Direct Reinforcement learning, stochastic games and computational finance. Moody is advising OGI Computer Science Ph.D. student Matthew Saffell in the areas of time series analysis and computational finance. Matt holds previous degrees in mathematics and computer science. The fourth member of Moody’s group is Caltech alum Aron Rempel, a research programmer with industrial experience at Intel and Applied Materials and who holds a Master of Science in Computational Finance.

During Fall 2003, Moody’s group conducted research on computational finance, economic forecasting, stochastic games, and reinforcement learning (RL). One achievement during calendar year 2003 was the development of Stochastic Direct Reinforcement (SDR), an algorithm that shows promise of making RL tractable and able to solve real-world problems.

The development of practical RL algorithms has been an elusive, long-term goal of artificial intelligence research. In RL, agents seek to optimize sequences of inter-dependent decisions or actions over time through trial and error experience. Most RL research has focused on methods derived from dynamic programming. These methods based on Markov Decision Processes (MDPs) learn a value function (VF), and represent actions or policies only implicitly. VF algorithms are often found to be inefficient, to produce unstable solutions, and to have difficulty
scaling up to large problems. VF methods quickly encounter Bellman’s “curse of dimensionality”. In fact, only one application of RL has made it out of the laboratory and into the real world: IBM’s TD-Gammon backgammon game (bundled with the OS/2 Family Funpack).

In contrast, Direct Reinforcement (DR) methods (policy gradient and policy search) represent policies explicitly and do not require that a VF be learned. The SDR algorithm [44] generalizes the policy gradient work of [4] and the recurrent reinforcement learning (RRL) algorithm of [46] [45]. This stochastic algorithm is formulated for probabilistic actions, partially-observed states and non-Markovian policies. With SDR, an agent represents actions explicitly, does not need to learn a value function and can naturally incorporate recurrent structure that is intrinsic to many potential applications. Via use of recurrence, the short term effects of an agent’s actions on its environment can be captured, leading to the discovery of more effective policies. SDR computes utility gradients using a set of recursion equations that capture m-th order recurrence, and uses stochastic gradient ascent to optimize a stochastic policy.

Stochastic matrix games provide a microcosm for studying the properties of dynamic learning in multi-agent systems (MAS), as well as in more general dynamical contexts. We find that recurrence becomes important when interacting with predictive or reactive agents in a multi-agent system. One important goal of current MAS research is to devise mechanisms that lead to cooperation amongst multiple, autonomous learning agents. The iterated prisoners’ dilemma (IPD) is a classic example from game theory in which groups of agents can learn either to compete (defect) or to cooperate, and thereby achieve greater overall utility.

In IPD play with a dynamic tit-for-tat opponent, we find that non-recurrent SDR agents prefer the static, Nash equilibrium strategy of defection, while recurrent m=1 players are able to learn to cooperate, thus finding the globally-optimal cooperative equilibrium. Against a tit-for-two-tat opponent, an m=2 recurrent SDR player learns the optimal dynamic strategy of alternating cooperation and defection. In self-play, we find that non-recurrent SDR agents learn to always defect, while recurrent SDR players learn to always cooperate. For a multi-agent generalization of IPD in which an SDR learner plays a heterogeneous population of opponents, SDR is able to learn the tit-for-tat strategy. Hence, SDR agents can learn to encourage cooperation in their opponents.

These empirical results demonstrate the importance of recurrence and non-Markovian representations in reinforcement learning. The SDR algorithm enables much simpler representations and faster learning than VF-type reinforcement algorithms and may have potential to scale up to larger or harder problems. Recurrent SDR may find application in various dynamic contexts such as planning, decision support, robotics, e-commerce, auctions, communication networks, distributed computing or other cooperative multi-agent systems.

3.4 Combinatorial Algorithms

3.4.1 Edge Modification Problems

In a clustering problem one has to partition a set of elements into homogeneous and well-separated subsets. From a graph theoretic point of view, a cluster graph is a vertex-disjoint union of cliques. The clustering problem is the task of making the fewest changes to the edge set of an input graph so that it becomes a cluster graph. We have studied the complexity of three variants of the problem. In the Cluster Completion variant edges can only be added. In Cluster Deletion, edges can only be deleted. In Cluster Editing, both edge additions and edge
Figure 2: Iterated Prisoner Dilemma with a Tit-for-Tat Opponent. Fractions of cooperation (C) and defection (D) for the SDR/TFT players as a function of time. For the match with a standard, non-recurrent SDR player (top panel), the dominant outcome D/D is established after seven trials. This is the non-cooperative Nash equilibrium. For the match with a recurrent SDR player (bottom panel), the SDR agent discovers the joint benefits of cooperation, and the dominant outcome of C/C is established after about eighty trials. This is the globally-optimal cooperative equilibrium. The discovery of cooperative behavior is important for multi-agent systems consisting of interacting autonomous learning agents.
deletions are allowed. We also studied these variants when the desired solution must contain a prespecified number of clusters. In [53] we showed that Cluster Editing is NP-complete, Cluster Deletion is NP-hard to approximate to within some constant factor, and Cluster Completion is polynomial. When the desired solution must contain exactly $p$ clusters, we showed that Cluster Editing is NP-complete for every $p \geq 2$; Cluster Deletion is polynomial for $p = 2$ but NP-complete for $p > 2$; and Cluster Completion is polynomial for any $p$. We also gave a constant factor approximation algorithm for a variant of Cluster Editing when $p = 2$.

In [24] we present efficient fixed-parameter algorithms for the NP-complete edge modification problems CLUSTER EDITING and CLUSTER DELETION. Here, the goal is to make the fewest changes to the edge set of an input graph such that the new graph is a vertex-disjoint union of cliques. Allowing up to $k$ edge additions and deletions (CLUSTER EDITING), we solve this problem in $O(2.27^k + |V|^3)$ time; allowing only up to $k$ edge deletions (CLUSTER DELETION), we solve this problem in $O(1.77^k + |V|^3)$ time. The key ingredients of our algorithms are two easy to implement bounded search tree algorithms and an efficient polynomial-time reduction to a problem kernel of size $O(k^3)$. This improves and complements previous work. Finally, we discuss further improvements on search tree sizes using computer-generated case distinctions.

In [25] we present a framework for an automated generation of exact search tree algorithms for NP-hard problems. The purpose of our approach is two-fold—rapid development and improved upper bounds. Many search tree algorithms for various problems in the literature are based on complicated case distinctions. Our approach may lead to a much simpler process of developing and analyzing these algorithms. Moreover, using the sheer computing power of machines it may also lead to improved upper bounds on search tree sizes (i.e., faster exact solving algorithms) in comparison with previously developed “hand-made” search trees. Among others, such an example is given with the NP-complete CLUSTER EDITING problem (also known as CORRELATION CLUSTERING on complete unweighted graphs), which asks for the minimum number of edge additions and deletions to create a graph which is a disjoint union of cliques. The hand-made search tree for CLUSTER EDITING had worst-case size $O(2.27^k)$, which now is improved to $O(1.92^k)$ due to our new method. (Herein, $k$ denotes the number of edge modifications allowed.)

**Dynamic Graph Algorithms:** The problem of dynamically recognizing a graph property calls for efficiently deciding if an input graph satisfies the property under repeated modifications to its set of vertices and edges. The input to the problem consists of a series of modifications to be performed on the graph. The objective is to maintain a representation of the graph as long as the property holds, and to detect when it ceases to hold. We have solved the dynamic recognition problem for the class of cographs and some of its subclasses [51]. Our approach is based on maintaining the modular decomposition tree of the dynamic graph, and using this tree for the recognition. We give the first fully dynamic algorithm for maintaining the modular decomposition tree of a cograph. We thereby obtain fully dynamic algorithms for the recognition of cographs, threshold graphs, and trivially perfect graphs. All these algorithms work in constant time per edge modification and $O(d)$ time per $d$-degree vertex modification.
3.4.2 The Stacker Crane Problem

Given an edge-weighted transportation network \( G \) and a list of transportation requests \( L \), the **Stacker Crane Problem** is to find a minimum-cost tour for a server along the edges of \( G \) that serves all requests. The server has capacity one, and starts and stops at the same vertex. In this paper, we consider the case that the transportation network \( G \) is a tree, and that the requests are chosen randomly according to a certain class of probability distributions. We show that a polynomial time algorithm by Frederickson and Guan (Nonpreemptive ensemble motion planning on a tree, Journal of Algorithms, 15 (1993), pp. 29–60), which guarantees a \( 4/3 \)-approximation in the worst case, on almost all inputs finds a minimum-cost tour, along with a certificate of the optimality of its output.

3.4.3 Arc-Annotated Sequences

In [3] we present exact algorithms for the NP-complete **LONGEST COMMON SUBSEQUENCE** problem for sequences with nested arc annotations, a problem occurring in structure comparison of RNA. Given two sequences of length at most \( n \) and nested arc structure, one of our algorithms determines (if existent) in \( O(3.31^{k_1+k_2} \cdot n) \) time an arc-preserving subsequence of both sequences, which can be obtained by deleting (together with corresponding arcs) \( k_1 \) letters from the first and \( k_2 \) letters from the second sequence. A second algorithm shows that (in case of a four letter alphabet) we can find a length \( l \) arc-annotated subsequence in \( O(12^l \cdot \ln n) \) time. This means that the problem is fixed-parameter tractable when parameterized by the number of deletions as well as when parameterized by the subsequence length. Our findings complement known approximation results which give a quadratic time factor-2 approximation for the general and polynomial time approximation schemes for restricted versions of the problem. In addition, we obtain further fixed-parameter tractability results for these restricted versions.

In [3] we start a study of pattern matching for arc-annotated sequences. An \( O(nm) \) time algorithm is given for the problem to determine whether a length \( m \) sequence with nested arc annotation is an arc-preserving subsequence (aps) of a length \( n \) sequence with nested arc annotation, called APS(NESTED,NESTED). Arc-annotated sequences and, in particular, those with nested arc structure are motivated by applications in RNA structure comparison. Our algorithm generalizes results for ordered tree inclusion problems and it can be used to accelerate a recent fixed-parameter algorithm for LAPCS(NESTED,NESTED), which is the problem of computing a longest common aps of two sequences with nested arc annotations. In particular, the presented dynamic programming methodology implies a quadratic-time algorithm for an open problem posed by Vialette.

Contact maps are a central model to capture the structure of biological molecules, e.g., proteins. Looking for structural features of a molecule leads in this model to a structural version of the classical pattern matching problem, which, in general, turns out to be NP-hard. Answering an open question by Vialette, we show how to solve the pattern matching problem for a restricted class of contact map patterns in polynomial time [22].

3.4.4 Fixed-parameter tractability and approximation.

A central question in computational biology is the design of genetic markers to distinguish between two given sets of (DNA) sequences. This question is formalized as the NP-complete **DISTINGUISHING SUBSTRING SELECTION** problem (DSSS for short) which asks, given a set
of "good" strings and a set of "bad" strings, for a solution string which is, with respect to Hamming metric, "away" from the good strings and "close" to the bad strings. More precisely, given integers $d_g$, $d_b$, and $L$, we ask for a length-$L$ string $s$ such that $s$ has Hamming distance at least $d_g$ to every length-$L$ substring of the good strings and such that every bad string has a length-$L$ substring with Hamming distance at most $d_b$ to $s$.

Studying the parameterized complexity of DSSS, we show in [26] that DSSS is W[1]-hard with respect to its natural parameters. This, in particular, implies that a recently given polynomial-time approximation scheme (PTAS) by Deng et al. cannot be replaced by a so-called efficient polynomial-time approximation scheme (EPTAS) unless an unlikely collapse in parameterized complexity theory occurs. This is seemingly the first computational biology problem for which such a border between PTAS (which exists) and EPTAS (which is unlikely to exist) could be established.

By way of contrast, for a special case of DSSS, we present an exact fixed-parameter algorithm solving the problem efficiently. In this way, we also exhibit a sharp border between fixed-parameter tractability and intractability results.

In [63] we initiate a systematic study of the ROW DELETION($B$) problem on matrices: For a fixed "forbidden submatrix" $B$, the question is, given an input matrix $A$ (both $A$ and $B$ have entries chosen from a finite-size alphabet), to remove a minimum number of rows such that $A$ has no submatrix which is equivalent to a row or column permutation of $B$. An application of this question can be found, e.g., in the construction of perfect phylogenies. Establishing a strong connection to variants of the NP-complete HITTING SET problem, we show that for most matrices $B$ ROW DELETION($B$) is NP-complete. On the positive side, the relation with HITTING SET problems yields constant-factor approximation algorithms and fixed-parameter tractability results.

3.5 Networking

Several algorithmic developments in networking are described in the report of the ICIR Group.

3.6 Computational Complexity

3.6.1 Reducibility to Sets with Low Information Content

The paper [47] studies for various natural problems in NP whether they can be reduced to sets with low information content, such as branches, P-selective sets, and membership comparable sets. The problems that are studied include the satisfiability problem, the graph automorphism problem, the undirected graph accessibility problem, the determinant function, and all logspace self-reducible languages. Some of these are complete for complexity classes within NP, but for others an exact complexity theoretic characterization is not known. Reducibility of these problems is studied in a general framework introduced in this paper: prover-verifier protocols with low-complexity provers. It is shown that all these natural problems indeed have such protocols. This fact is used to show, for certain reduction types, that these problems are not reducible to sets with low information content unless their complexity is much less than what it is currently believed to be. The general framework is also used to obtain a new characterization of the complexity class $L$: it is the class of all logspace self-reducible sets in $L^{L_{redu}}$. 
3.6.2 CVerbose Languages

A language is called \( m, n \)-Verbose if there exists a Turing machine that enumerates for any \( n \) words at most \( m \) possibilities for their characteristic string. We compare this notion to \( m, n \)-fa-Verbose, where instead of a Turing machine a finite automaton is used. Using a new structural diagonalisation method, where finite automata trick Turing machines, we prove that all \( m, n \)-Verbose languages are \( h, k \)-Verbose, if all \( m, n \)-fa-Verbose languages are \( h, k \)-fa-Verbose. In other words, Turing machines and finite automata behave in exactly the same way with respect to inclusion of Verbose classes. This identical behaviour implies that the Nonseup Theorem also holds for finite automata. As an application of the theoretical framework, we prove a lower bound on the number of bits needed to be communicated to finite automata protocol checkers for nonregular protocols.

3.6.3 Reachability in Digraphs with Low Independence Number

How difficult is it to find a path between two vertices in finite directed graphs whose independence number is bounded by some constant \( k \)? The independence number of a graph is the largest number of vertices that can be picked such that there is no edge between any two of them. The complexity of this problem depends on the exact question we ask: Do we only wish to tell whether a path exists? Do we also wish to construct such a path? Are we required to construct the shortest path? Concerning the first question, it is known that the reachability problem is first-order definable for all \( k \). In contrast, the corresponding reachability problems for many other types of finite graphs, including dags and trees, are not first-order definable. Concerning the second question, in [62] it is shown that we cannot only construct paths in logarithmic space, but that there even exists a logspace approximation scheme for constructing them. In contrast, for directed graphs, undirected graphs, and dags we cannot construct paths in logarithmic space (let alone approximate the shortest one), unless complexity class collapses occur. Concerning the third question, it is shown that even telling whether the shortest path has a certain length is NL-complete and thus as difficult as for arbitrary directed graphs.

3.7 Research on Combinatorial Games

Combinatorial Game Theory is the study of combinations of two-person games with perfect information. Examples of such games include Amazons, Chess, Dominering, and Go. The modern form of the theory began in the early 1970s. One early result was Conway's discovery that any such game, \( G \), has a canonical form, which encapsulates all the information needed to determine who wins when \( G \) is played as a sum with other combinatorial games.

Twenty-five years ago, most applications of Combinatorial Game Theory were restricted to idealized mathematical games and puzzles, and only a few "real" games, such as Dots-and-Boxes. More recently, the theory has found applications to many more real-world games, such as Amazons, Fox-and-Geese, and the Hawaiian game Konane. With some modifications, the theory has also been applied successfully to endgames in Go and, to a lesser extent, Chess.

At ICSI in 2003, there were two projects in Combinatorial Games:
3.7.1 Quantitative Go

Work reported last year on Quantitative Go continued in 2003, with the active participation of Elwyn Berlekamp, Brian Carr, and Bill Spight. The long-term goal is the development of software which would enable skilled human users to analyze real professional Go endgames and find fatal mistakes. We believe this will provide a powerful demonstration of the power and relevance of combinatorial game theory, and that it will lead to more fruitful interactions between professional Go players, computer Go buffs, and combinatorial game theorists.

The goal is to start with a record of a professional game, go to the terminal position, and begin working backward to earlier stages of the game. As each new move is taken back, local alternative lines of play are entered. The main thesis, evidenced by prior work such as [59], is that good human players are quite good at narrowing the search down to a very small number of promising regional moves, but they are much less adept at comparing the values of different moves in different regions of the board. Under idealized circumstances (including total independence of regions), a theory called thermography reduces the latter problem to calculations which are fast for computers but too cumbersome to be usable by competitive human Go players. Extended thermography developed in the past decade has given us reason to hope that our goals may be achievable sometime within the next several years.

The biggest step forward in 2003 was the preliminary implementation of a program called “Automerge”, which assists the human analyzer in his efforts to partition the overall board position into regions. Preliminary techniques have been devised for analyzing “federations” of regions which are weakly dependent on each other. Future efforts will include studies and eventual implementations of a “dogmatic” theory to provide a more general method of handling certain types of complicated loopy positions called kos.

3.7.2 Cgsuite

Since canonical forms are essentially algebraic and are generally cumbersome to calculate by hand, Combinatorial Game Theory lends itself well to symbolic computational analysis, and computer applications have grown in importance in recent years. In many cases, the solution to a problem becomes clear only after a large number of special cases have been calculated; other problems are inherently computationally hard.

In the fall of 2002 Aaron Siegel began work on an open-source software program, Combinatorial Game Suite (cgsuite). Cgsuite calculates canonical forms for a wide variety of games. It can perform many algebraic operations on canonical forms. Cgsuite is the first program to do combinatorial game analysis, but it offers several advantages over previous software, including support for several kinds of loopy games. In the spring of 2003, Berlekamp and Siegel used this program to essentially solve a loopy English children’s game called “Fox-and-Geese” for boards of size Nx8. This entailed both an asymptotic analysis of the game and many improvements to cgsuite: developing new algorithms, optimizing existing ones, and improving the overall stability of the software. A copy of cgsuite can be obtained at http://cgsuite.sourceforge.net; source code and documentation are also posted there. In August 2003, this program was field-tested and enthusiastically received by a group of undergraduate mathematics teachers who attended an MAA Workshop on Combinatorial Games held at Gustavus University in St. Peter, MN.

Although cgsuite can perform a broad set of computational tasks in Combinatorial Game
Theory, it still has several limitations. In some cases, one is interested only in who wins a particular game, and less interested in the exact canonical form. Software geared towards simply determining the winner can be made far more efficient than cgsuite; for example, 10 x 10 Domineering is known to be a first player win, but the exact canonical form of even 7 x 7 Domineering seems well beyond the reach of present computational resources. Also, quantitative Go includes some specialized theories that cgsuite is not equipped to handle.

It is known that canonical forms contain much information that is almost always irrelevant to the play of a game. Berlekamp [9] observed that by adding an additional idempotent component to a game, one can often simplify the play in a way that discards this extraneous information. (The nature of the idempotent component depends on which game is being analyzed; the appropriate component for one game might destroy too much or too little information about another). Theoretical research efforts at ICSI in 2003 included a search for a “generalized reduced canonical form” which would extend the applicability of these results.

References


4 Artificial Intelligence and its Applications

The Artificial Intelligence group continues its long term study of language, learning, and connectionist neural modeling. The scientific goal of this effort is to understand how people learn and use language. The applied goal is to develop systems that support human-centered computing through natural language and other intelligent systems. Several shorter term goals and accomplishments are described in this report. There is continuing close cooperation with other groups at ICSI, at UC Berkeley, and with external sponsors and other partners. There are three articulating subgroups and this report summarizes their work.

4.1 The Neural Theory of Language, J. Feldman and S. Narayanan

It has long been known that people would prefer to talk with computer systems in natural language if they could. The problem of communicating with machines is becoming increasingly important to society because computers will soon be embedded in nearly every artifact in our environment. But how easy will it be for people of all ages and abilities to use them? In ten years or less, virtually every device in our environment will have a computer in it. This raises the specter of an embedded computing malaise—every device will have its own interface that the user has to learn. In the world of embedded computing, there could be thousands of idiosyncratic interfaces to learn. Natural interaction with computerized devices and systems requires a conceptual framework that can communicate about requests specified in ordinary language. Systems may well need to tell their human users what is going on, ask for their advice about what to do, suggest possible courses of action, and so on. The central goal of this project is to provide a conceptual basis and a linguistic framework that is rich enough to support a natural mode of communication for this evolving human/machine society.

While the usefulness of natural language usage (NLU) systems has never been questioned, there have been mixed opinions about their feasibility. Most current research is focused on goals that are valuable, but fall far short of what is needed for the natural interactions outlined above. We believe that recent advances in several areas of linguistics and computational theory and practice now allow for the construction of programs that will allow robust and flexible integrated language interaction within restricted domains.

For many years, Jerome Feldman has studied various connectionist computational models of conceptual memory and of language learning and use. George Lakoff and Eve Sweetser have worked on the relation between linguistic form, conceptual meaning, and embodied experience. Over the past dozen years, the group has explored biologically plausible models of early language learning [5] and of embodied metaphorical reasoning [16]. About two years ago, we extended our efforts on modeling child language acquisition from individual words and phrases [21, 4] to complete utterances. This required us to develop a formal notion of what it means to learn the relationship between form and meaning for complete sentences. Many groups, including ours [8] have worked on algorithms for learning abstract syntax, but we decided that it was time to look directly at learning form-meaning pairs, generally known as constructions.

After an intensive effort by the whole research group, we now have an adequate formalization of constructions and are moving ahead with the project of modeling how children learn grammar from experience. This will form the core of a dissertation by Nancy Chang, a UCB doctoral student. But we also realized that our formalized notion of linguistic constructions
that systematically links form to conceptual meaning is potentially a breakthrough in achieving robust and flexible NLU systems.

The most novel computational feature of the NTL effort is the representation of actions: executing schemas (x-schemas), so named to remind us that they are intended to execute when invoked. We represent x-schemas using an extension of a computational formalism known as Petri nets [15]. As discussed below, x-schemas cleanly capture sequentiality, concurrency, and event-based asynchronous control. With our extensions they also model the hierarchy and parameterization needed for action semantics.

Our goal is to demonstrate the unification of two powerful linguistic theories, embodied semantics and construction grammar, together with improvement in HCI based on NLU. Over the last year we explored extending our existing pilot system to moderate sized applications in real HCI settings and develop the methodology needed for large scale realization of NLU interaction. This involves formalization and additional research in cognitive linguistics, development of probabilistic best fit algorithms and significant system integration. Much of the group’s effort over the past year has gone into developing these formalisms and producing a pilot version of the integrated language understanding system.

For concreteness, we have chosen a specific task domain for the proof-of-concept demonstration of our research. We are constructing a system for understanding and responding to dialog with tourists, initially focused on Heidelberg, Germany. This applied project is being carried out in cooperation with a partner group at EML in Heidelberg, which has built an extensive data base describing their city (www.villa-bosch.de/english/research) and will implement the detailed actions for using it based on our natural language analysis. This cooperation will bring several benefits to the project and provides clear milestones for evaluating our effort. This project (called EDU for Even Deeper Understanding) has been in operation since July 2000, with multi-year funding from the Klaus Tschira Foundation. Robert Porzel, of EML, joined our group for the calendar year 2001. John Bryant from ICSI spent the last half of 2002 working at EML in Heidelberg. A major effort of this collaboration was the first international workshop on Scalable Natural Language Understanding Systems (Scanalu), held in Heidelberg. A second Scanalu workshop will be held in Boston in May, 2004.

This effort is also closely linked to the SmartKom project, which is discussed in the Speech section of this annual report. Another cooperation between the Speech and Language groups is the human interface section of the CITRIS proposal to the state of California, funded this year. CITRIS is a large multi-disciplinary effort that has many subprograms, one of which is a new group at ICSI, the Berkeley Center for the Information Society (BCIS). BCIS is just completing its first year; a brief summary is included as part of this report.

In 2002 the NTL group began working on another large cooperative NLU project, after winning in a competitive grant competition in the Aquaint program of the U.S. Defense Department ARDA organization. The group teamed with Prof. Marti Hearst (SIMS, UCB) and Prof. Chris Manning (Stanford) to study deep inferencing techniques and corpus based techniques for deriving the conceptual semantics needed to achieve this. Our effort is being integrated into an ambitious overall program to significantly advance the automated analysis of information. This makes significant use of our basic work on both grammar and inference and is also contributing to it. In 2003, the NTL group demonstrated several of its results at various ARDA workshops and meetings. The group was also successful in attaining funding for Aquaint Phase 2, in collaboration with Stanford and U. Texas at Dallas.
The core NTL computational question is finding the best match of constructions to an utterance in linguistic and conceptual context. One of the attractions of traditional phrase structure grammars is the fact that the time to analyze (parse) a sentence is cubic in the size of the input. If one looks at the comparable problem for our more general construction grammars, context-free parsing becomes NP complete (exponential) in the size of the input sentence and thus impractical. But people do use larger constructions to analyze language and we believe that we have two insights that seem to render the problem of construction analysis tractable. The general computational point is that our task of finding a best-fit analysis and approximate answers that are not always correct presents a more tractable domain than exact symbolic matching. More importantly, our integrated constructions are decidedly not context-free or purely syntactic. We believe that constraints from both semantics and context will be sufficiently constraining that it will be possible in practice to build best-fit construction matchers of the required scale. John Bryant, a CS doctoral student, has completed a Masters thesis on this topic and is planning to continue for his doctorate.

This sequence of operations: surface analysis, construction parse, SemSpec, simulation and inference is repeated for every clause. The current pilot system does not make use of extensive context or world knowledge, but these are central to our new design. There is currently a great deal of renewed effort to develop ontologies of words and concepts for a wide range of semantic domains (Fikes 1994). After analyzing these efforts, we have decided against committing to any one of the competing formulations and have instead defined an Application Programming Interface (API) that our system can use to access information from any source. A preliminary version of this is used in the pilot system and we will evolve the API as experience requires. The current API has the usual commands for adding information and some special ones for retrieving ordered lists of concepts most likely to fulfill a request. This also facilitates our interaction with the EML project (EDU) and the German SmartKom effort.

There was also a significant effort on related problems that elucidate or exploit our main results. Ben Bergen completed a UCB linguistics thesis using a statistical, corpus-based approach in combination with psycholinguistic experimentation, to explore probabilistic relations between phonology on the one hand and syntax, semantics, and social knowledge on the other. The group has developed a formal notation for Embodied Construction Grammar (ECG), which plays a crucial role in a larger, simulation-based language understanding system. There was a full day session on ECG at the 2003 International Conference on Cognitive Linguistics in Spain.

We also devised an experimental means by which to test the psychological reality of construal, the variable, context-specific understanding of the semantic pole of linguistic constructions. A paper on this was presented at the 2003 Cognitive Science Conference.

Nancy Chang and others have continued developing representations and algorithms useful for an embodied approach to language acquisition and use. She worked with colleagues to flesh out different aspects of a simulation-based approach to language understanding, including a formal representation for linguistic constructions. A version of the formalism is incorporated into her thesis research, which focuses on the development of an algorithm that learns such constructions from a set of utterance-situation pairs.

In 2003 there was a very significant increase in the use of the group’s results in UCB courses and in linguistics research. Collaboration with the FrameNet project has been broadened and deepened with positive results for both efforts, some of which are described in this report. Eve Sweetser and Jerome Feldman ran an interdisciplinary graduate seminar in Fall 2002 and several of the research efforts from that class are being incorporated into the project. Several new UCB
doctoral students have become involved with the group including John Bryant, Ellen Dodge, Olya Gurevich, Behrang Mohii, Eva Mok, Shweta Narayan, and Steven Sinha.

Thus the NTL group has, over the last year, formalized and significantly extended its work on language learning and use based on deep conceptual semantics. Both the learning sub-task and the performance HCI system are moving ahead in collaboration with other efforts at ICSi and elsewhere. One of these involves a large effort on Modeling Services on the Semantic Web. Another involves an expressive probabilistic model of inference. A third involves applications of FrameNet for Natural Language Processing.

4.1.1 Semantic Web Services

The Semantic Web is an exciting vision for the evolution of the World Wide Web. Adding semantics enables structured information to be interpreted unambiguously. Precise interpretation is a necessary prerequisite for automatic Web search, discovery and use. Services are a particularly important component of the Semantic Web. A semantic service description language can enable a qualitative advance in the quality and quantity of e-commerce transactions on the Web. The DAML Services Coalition, under the guise of DAML-S, has taken some important first steps in this direction. The model of actions, processes and events developed within the NTL project provides a natural, distributed operational semantics that may be used for simulation, validation, verification, automated composition and enactment of DAML-S-described Web services. The details of our approach are described in the paper “Analysis and Simulation of Web Services”, which was the third most downloaded paper from the journal Computer Networks for the year 2003. The benefits of our approach include: 1) Formal executable semantics: a service description is fully represented using the machinery of situation calculus and its execution behavior unambiguously described using Petri Nets. 2) Analysis techniques and tools: mapping DAML-S onto situation calculus and Petri Nets allows us to tap into a rich repository of analysis techniques and tools. 3) Service implementation tool: we mapped the DAML-S service description to an existing process model which was able to perform simulation, enactment and analysis of composite service descriptions. 4) Complexity and reasoning: the expressive power of the DAML-S process model compares to ordinary Petri Nets. We identified more tractable subsets of DAML-S which trade expressiveness for more efficient analysis for verification, composition and model checking.

4.1.2 Applications of FrameNet

As a rich, theoretically well founded lexical semantic resource, FrameNet can potentially benefit a variety of Natural Language Processing systems. In 2003, our goal was to make FrameNet (http://www.icsi.berkeley.edu/framenet) usable by making the resource available to researchers in a variety of formats and by demonstrating its utility for applications in Natural Language Processing. The following specific projects were accomplished to address this goal.

FrameNet meets the semantic web (http://www.daml.org), a widely used language related to the Semantic Web initiative (http://www.semanticWeb.org). The DAML language is being developed as an extension to XML and the Resource Description Framework (RDF). The latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. We developed an automatic translator from FrameNet to DAML+OIL.
which is being updated to reflect FrameNet2 data. With periodic updates as the FrameNet data increases, we expect it to become useful for various applications on the Semantic Web. This effort was reported at the International Conference on the Semantic Web.

FrameNet for Information Extraction Behrang Mohit and Srin Narayanan have developed an open domain Information Extraction (IE) algorithm that learns domain specific extraction rules using seed patterns derived from FrameNet annotations. The initial pattern set is obtained by automatically translating FrameNet annotations for a particular target frame into extraction patterns covering that frame. With just the initial annotations, we were able to obtain a precision score of 70%. We then added information from the WordNet database, which offers a different type of hierarchical results for a given term. We used the Hypernym relations in WordNet (that are similar to ISA hierarchical relation) to expand our lexicon by automatically identifying the right WordNet node(s) for each of the frame elements to achieve an optimum f-score. With this addition, we were able to achieve an f-score of around .7 (comparable to many hand-built IE systems). This effort was reported at the HLT-NAACL conference in 2003.

4.1.3 CPRM: A Scalable and Expressive Model of Actions and Events

![Diagram of Probabilistic Models and Inference Space]

Figure 3: Probabilistic Models and Inference Space

Reasoning about structured stochastic dynamic systems requires modeling coordinated temporal processes and complex, structured states. A significant amount of work has gone into different aspects of the overall problem.¹ Figure 3 maps out the space of relevant probabilistic modeling and inference techniques along three basic dimensions (extended from the description in [1]). The dimension along the x-axis (left-right) depicts the increasing expressiveness of the action model going from no temporal dimension to linear temporal models (such as Markov Chains and HMMs to branching and coordinated models such as Petri Nets). The y-axis (vertical going up) corresponds to increasing the complexity of the state representation (from flat attribute-value structures to modular structures like factor graphs or Bayes Nets) where the joint probability is computed recursively as a product of conditional distributions. The z-axis (into the plane) corresponds to increasing the richness of the overall representation (from propositional

¹In all these cases, we can have continuous variables as well as discrete ones. For the purposes of this exposition, all the comments here apply to both types of states and actions.
to relational). The origin of the space is an unstructured probabilistic state vector representation with no explicit temporal or relational information.

To model complex domains with rich event structures for QA, we have developed a model (Narayanan and Feldman 2004) that combines the modular and structured probabilistic framework of PRMs with the flexible control and coordinated dynamics of extended Petri nets. We call this model Coordinated Probabilistic Relational Models or CPRM. These can be seen as extending an ongoing community effort to add more sophisticated temporal and control capabilities to probabilistic inference models. We have implemented the CPRM model of inference and are currently testing it for use in Question Answering as part of the ARDA sponsored AQUAINT program in collaboration with Stanford University and University of Texas, Dallas.

4.2 FrameNet, C. Fillmore

The NSF-sponsored FrameNet project began in 1997 with NSF IRI-9618838 “Tools for Lexicon Building” with the goal of creating an online lexicon for English, based on frame semantics and supported by corpus evidence. The project is now in its second major phase, having received 2.1M in the year 2000 (NSF HCI-0086132, “FrameNet++: An Online Lexical Semantic Resource and its Application to Speech and Language Understanding”) for expanding the lexical database itself and for pilot projects on a battery of NLP applications that make use of it. Applications under study include automatic word-sense disambiguation, automatic semantic role labeling, machine translation, information extraction, question answering, and text understanding.

In both phases, the main task is to document from actual text data the varieties of uses of English lexical items. Each meaning of each word is associated with a semantic frame which represents the conceptual structure that underlies it. The frame contains a set of frame elements, which are frame-specific names and definitions for the participants and props involved in the situation described by the frame. Sentences that exemplify each word in a frame are automatically extracted from the corpus and then manually annotated to show which parts of the sentence represent which frame elements. These annotated sentences are then automatically categorized and combined with the manually written definitions to produce the lexical entries for the words in each frame, demonstrating all the syntactic patterns in which it can occur.

Thus the FrameNet Database provides (1) a collection of semantically annotated examples for each sense of each word, (2) links to descriptions of the conceptual structures (the semantic frames) which underlie each such sense, and (2) details of the ways in which the semantic roles (frame elements) in each frame are syntactically realized in sentences containing the word, both individually and in combinations.

The corpus on which these observations were based in the first phase was the British National Corpus (100M running words); we have now added an American Newspaper Corpus made available through the Linguistic Data Consortium (University of Pennsylvania), and we are actively participating in the development of the new American National Corpus, headed by Prof. Nancy Ide of Vassar College.

The ICSI FrameNet Project has continued the development of its uniquely detailed lexicon and has added a number of enhancements of the database, while cooperating in various multilingual expansions and shaping the role of the database in various NLP applications.

The first, preliminary data release from the second phase of the project took place in October, 2002, and contained more than 6,800 lexical units in approximately 400 frames; the data
is both accessible for viewing on the redesigned project website and available to researchers for downloading, in HTML and XML formats. Requests for downloading arrive almost daily, from individual researchers and from research institutions around the world, providing growing evidence of its usefulness in areas of language processing, language pedagogy and linguistic research.

Projects are underway to build databases like FrameNet for limited domains in Spanish, German, and Japanese. The Spanish and German projects result partly from current and past support to ICSI visitors, through funding from Spain and Germany, but the Japanese development was independent. Also, ICSI and TEKES are sponsoring an effort based in Tampere, Finland, to create FrameNet-style lexical descriptions, for use in semantic analysis of Finnish social services documents for an automated help system.

A major revision of the software we use for defining frames and annotating sentences is just being completed. The new software will be cleaner, more modular, and able to handle the new types of semantic information we are now adding to the database, including a wide variety of relations among frames and frame elements, and a variety of semantic types. The new software can also be distributed to collaborators building similar databases, either for other languages or in specialized domains.

FrameNet published a collection of articles in the International Journal of Lexicography – an issue dedicated to the project [2, 3, 6, 11, 12].

The new FrameNet data were released in January, 2004.

The FrameNet team put on a two-day workshop on December 8-9, 2003 sponsored by DARPA-IPTO, to present the procedures, accomplishments and possibilities of FrameNet to the general research community. Presentations by foreign collaborators were included, by means of recorded videotape from Barcelona and a live videoconference with Saarbruecken.

Attendees from academic, corporate and government research centers were:

Danny Bobrow (XEROX)
Ronald Brachman (DARPA)
Murray Burke (DARPA, emeritus)
Christiane Fellbaum (Princeton, WordNet)
Kenneth Forbus (NWU)
Dave Gunning (DARPA)
David Israel (SRI)
Doug Lenat (CYC)
Christopher Manning (Stanford)
George Miller (Princeton, WordNet)
John Sowa (IBM, emeritus)
Michael Witbrock (CYC)

ICSI participants included:
Jerome Feldman
Lokendra Shastri
Collin Baker (presenter)
Charles Fillmore (presenter)
Michael Ellsworth (presenter)
Josef Ruppenhofer (presenter)
Miriam Petrucc
4.2.1 Collaborations Initiated or Continued

- DARPA funding has been provided to have us consider aligning FrameNet data with other lexical and ontological resources, in particular WordNet and CYC.

- NSF subcontract with D. Jurafsky and M. Palmer for adding FrameNet-style semantic annotations to the PropBank corpus.

- TRANSCOOP (Alexander von Humboldt Foundation) travel grant has been awarded to support collaborative work between FrameNet and the SALSA project in Saarbruecken.

- ICSI international arrangements allow continued participation of foreign researchers in FrameNet activities (Germany - Petra Steiner; Spain - Carlos Subirats, Sira Palazuelos; Switzerland - Vincenzo Pallotta).

On Users of FrameNet data and their interests

There are two sites currently using the full set of FN software, including the database and the FNDesktop tool, one in Spain and one at MIT, and other sites plan to install it soon, in Texas and in Germany.

To date, the following areas of research have been identified by those who have downloaded the FN data. The number of users is given in parentheses. (Many claim more than one use.)

- Natural Language Understanding (32)
- Semantic Parsing (24)
- Word Sense Disambiguation (23)
- Information Extraction (33)
- Question Answering (25)
- Machine Translation (18)
- Research in Lexical Semantics (23)
• Lexicography (19)
• Classroom teaching of lexical semantics (9)

An independent browser for FrameNet data has been built in Japan by Professor Hiroaki Sato of Senshu University, "FrameSQL".

One of the panels in a competition called SENSEVAL has chosen to use FrameNet data as a training corpus for automatic semantic parsing; 36 participants have been announced. (http://www.clres.com/SensSemRoles.html)

4.3 Connectionist Models of Brain Function, L. Shastri and S. Narayanan

SHRUTI is a structured connectionist model of reflexive reasoning and decision making. The model can represent and process beliefs and utilities to make predictions, seek explanations, and identify actions that could make the world state more desirable. If the predictions and explanations drawn by the system suggest that undesirable states are imminent, the system automatically identifies actions that could prevent this from happening. In general, the system attempts to identify actions that would maximize the expected future utility.

Work on the SHRUTI model has demonstrated that a single causal structure, expressed as a neurally plausible network, can serve three purposes: (i) understand the world, (ii) predict the future, and (iii) plan for a better future.

SHRUTI was applied to model critical thinking under a project funded by the Army Research Institute. In consultation with the primary contractor (CTI), a well-known tactical game scenario was chosen for simulation. A SHRUTI knowledge base to support reasoning and decision-making in this scenario was developed, and several simulations were carried out.

4.3.1 Modeling cultural differences

The feasibility of using SHRUTI to model cultural differences was studied by encoding and simulating different scenarios to illustrate how SHRUTI could model cultural differences in reasoning and decision-making.

One scenario modeled two aspects of cultural differences: (i) how individuals from two cultures might differ in the utility (value) they associate with being punctual and (ii) how individuals from two cultures might differ in their sensitivity to social actions such as a guest leaving a party early. Two cultures, one individualistic and with a crisp conception of time, and another collectivist and with a more relaxed conception of time, were considered. The impact of different conceptions of time was illustrated with the use of different utilities associated with the "onTime" predicate. Being on time was modeled as much more rewarding in a crisp time culture than in a relaxed time culture. Differences between individualistic and collectivist cultures were realized as different probabilities, encoded in culture-specific rules, that leaving the party early will offend the party's host. As a result of these cultural differences, a crisp-time individualistic entity was inclined to leave the party early in order to make it to the meeting on time, whereas the relaxed-time collectivist entity preferred to remain.
4.3.2 Cognitive control

An understanding of cognitive control is critical for understanding complex cognitive function. Connectionist solutions to a set of neurally motivated control mechanisms, including monitoring, filtering, selection, maintenance, organization, and manipulation were designed, and it was shown that these primitives can be used to realize more complex control processes such as sub-goal focus.

4.3.3 Learning structured representations

A connectionist mechanism for learning new concepts and causal rules in SHRUTI was proposed. The mechanism combines ideas from recruitment learning and reinforcement learning.

4.3.4 A biologically realistic model of episodic memory

The hippocampal system (HS) consisting of the hippocampal formation and neighboring cortical areas in the ventromedial temporal lobe, plays a critical role in the encoding and retrieval of episodic memories. SMRITI (System for memorizing relational instances from transient impulses) is a computational model of episodic memory that demonstrates how a cortical activity representing an event or a situation can be transformed rapidly into a persistent and robust memory trace in the HS as a result of long-term potentiation.

The neural circuit required for encoding an episodic memory trace is fairly complex and idiosyncratic, but SMRITI shows that this complexity and idiosyncrasy is well matched by the complexity and idiosyncrasy of the HS architecture and local circuitry.

The model predicts the functional roles of each of the components of the HS and some of the cortical areas interacting with the HS, the properties of cortically expressed event schemas/frames underlying episodic memories, the sorts of memories that must persist in the HS for the long-term, the nature of memory consolidation, and memory deficits that would result from cell loss in the hippocampus and high-level cortical circuits encoding semantic knowledge.

Over the past year, two behavioral studies were designed based on predictions made by SMRITI. Behavioral experiments related to one of these studies relating to the fan-effect were begun in collaboration with psychologist Prof. G.A. Radavsky of the University of Notre Dame.

4.3.5 The Role of Cortico-SubCortical Loops in Planning and Working Memory: A Computational Model

Clinical and experimental research over the last decade has implicated neuroanatomic loops connecting the frontal cortex to the basal ganglia and thalamus in various aspects of planning and memory. There is by now robust evidence that the pre-frontal cortex plays a key role in various aspects of working memory and executive control. There is also clear evidence that the basal ganglia are closely involved with prefrontal cortex activity. From a functional viewpoint, while damage to the basal ganglia seems to produce cognitive deficits comparable to prefrontal cortex malfunction, teasing out the individual contributions has proven more problematic.

In 2003, Narayanan developed and implemented a computational model with the goal of fleshing out the role of cortical-basal-thalamic loops in planning and executive control. A distinguishing feature of the approach is a fine-grained model of basal-ganglia function that exploits
specific component connectivity and dynamics. The model is biologically plausible given current literature on the neurophysiology and disease pathology of the relevant brain regions. The model and preliminary results of applying the model to published behavioral data from Parkinson's (PD) and Huntington's (HD) subjects on a standard cognitive test (the Wisconsin Card Sorting Task (WCST)) are described in the paper "The role of cortico-basal-thalamic loops in cognition: a computational model and preliminary results", which appeared in the journal NeuroComputing.

This effort is ongoing and plans are under way to refine and test the model in two ways:

1. Design cognitive tests for which our models of planning, working memory and executive control are likely to predict non-obvious results.

2. Apply these tests on subjects with and without diseases affecting relevant brain regions (PD, HD) and evaluate the model with respect to the results.

References


5 Speech Processing

During 2003, the Speech Group continued to make progress in speech recognition and metadata extraction, but also expanded significantly into a full-fledged project on speaker recognition with new National Science Foundation support. The EARS program had its first full calendar year of effort, and the group received word that its meeting project would be bolstered by new European Union funding to begin in 2004. As usual, speech group research was varied, ranging from work on improved front-end processing of the speech signal to back-end analysis of structure in speech from meetings, and from highly speculative explorations to system prototyping.

The year’s Speech efforts were headed by continuing research staff members Jane Edwards, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg (ICSI and SRI), Andreas Stolcke (ICSI and SRI), Chuck Wooters, and Qifeng Zhu. Additionally, Nikki Mirghafori joined our staff in October. Dr. Mirghafori, formerly of Nuance Communications, received her PhD in our group in 1998. Our work also continued to be bolstered by external collaborators. In particular, Dan Ellis of Columbia University, Hyniek Hermansky (formerly of OGI and ICSI, now at IDIAP) and Herve Bourlard of IDIAP are all working with us on various projects described in this section. Other domestic and international colleagues have also played a critical role in our progress. Independent consultant George Doddington continues to work with the group to help formulate research directions and evaluation methods. As always, major contributions were also made by our team of students, research associates, postdoctoral Fellows, and international visitors. (see http://www.icsi.berkeley.edu/Speech/people.html for a current list of group members, collaborators, and alumni).

The sections below describe a number of the year’s major activities in speech processing. Consistent with our 2002 format we have organized the report in terms of major projects: EARS (large vocabulary speech recognition and related tasks); the Meeting Recorder project, whose meeting corpus was recently released publicly by the Linguistic Data Consortium (LDC); SmartKom, which focuses on machine query systems; and a range of projects that fall under the general heading of Acoustically Robust Systems. This last category includes continuing joint work with Infineon on exploring physiologically motivated algorithms, as well as studies of other tactics for robustness to noise and reverberation for application to spoken language recognition in meetings and to the SmartKom demonstrator. We also devote separate sections to the new work in speaker recognition, a brief update on our continuing study of matching speech algorithms to somewhat specialized computer architectures, and an introduction to a new project as part of a UC Berkeley team seeking to create inexpensive and robust devices to bring information technology to billions worldwide. While this listing includes many of our most significant projects, it is by no means exhaustive, but should provide a useful overview of the major activities in which we have been engaged this year.

5.1 EARS

As noted above, 2003 was the first complete year of sponsored research for the DARPA-sponsored “Effective, Affordable Reusable Speech-to-text” (EARS) program. The goal of this five-year project is to significantly advance the state of the art in multi-lingual speech recognition of both broadcast news and conversational telephone speech. The EARS program consists of two subprojects: Rich Transcription and Novel Approaches. ICSI is a team member on the Rich Transcription project (along with team leader SRI and partner University of Washington)
and the lead site for the Novel Approaches project (along with team members SRI, University of Washington, Columbia, and IDIAP in Switzerland; researchers from OGI, a partner in the original project, moved to IDIAP in 2003). Other teams involved in Rich Transcription projects for EARS are Cambridge University, IBM, and a BBN-led team that includes LIMSI, University of Pittsburgh, and the University of Washington. Lincoln Laboratory is developing speaker segmentation and tracking technology to support EARS, LDC is providing data, and NIST is under contract to handle the evaluation process. For the Novel Approaches endeavor, the other team outside of the ICSI-led group is a project at Microsoft. The following sections describe some of the major efforts for the year in these two projects.

5.1.1 Rich Transcription

The Rich Transcription team is working to generate more readable transcriptions of conversational and broadcast speech in multiple languages. "Readable" here means incorporating capitalization, punctuation, speaker labels, and other structural information implicit in the speech stream; but it also means making major improvements in core speech recognition performance, since word errors are still significant in this type of task. The team leader for the SRI+ICSI+UW Rich Transcription effort is Andreas Stolcke, who has a dual affiliation with ICSI and SRI. Barbara Peskin is the site leader for ICSI’s contributions to this project.

Testing in all cases is done on Broadcast News (BN) and Conversational Telephone Speech (CTS). The four major tasks comprising the SRI+ICSI+UW team’s Rich Transcription effort are: Core Automatic Speech Recognition (ASR) Algorithms; Rapid Development of ASR in New Languages and Domains (Portability); Metadata Extraction and Modeling; and Evaluation. We have been particularly focused on the first three, as SRI is responsible for the ultimate system integration.

In the area of Core ASR Algorithms, our primary responsibility is to work to incorporate innovations from our Novel Approaches effort into the Rich Transcription system. This year the Novel Approaches team achieved significant new results, and as a result we put a fair amount of effort into porting the new methods (described in a later section) to a state-of-the-art SRI speech recognition system. In addition to the transfer of Novel Approaches innovations, this year’s work in Core ASR also included a number of language modeling explorations, especially examining ways of using Latent Semantic Analysis and related techniques for topic-focused modeling, and preliminary investigations on speaker modeling for Broadcast News.

Our work on Portability has focused on two main efforts: developing a system for transcribing Broadcast News in Mandarin Chinese as part of the EARS goal of broadening our transcription technologies beyond English, and more speculative work in data-driven discovery of subword units that might provide a language-independent inventory of fundamental elements for building acoustic models. Constructing the Mandarin Broadcast News system was our biggest Portability challenge this year. While the system was modeled on SRI’s Broadcast News recognizer for English, this was our first foray into Mandarin recognition, requiring significant infrastructure work on dictionary development, data preparation, and accommodations to handle non-ASCII character sets. Further, in addition to the usual difficulties involved in recognizing broadcast data - the intermingling of speech and non-speech events, the mix of speakers and speaking styles, and the varied channel conditions and acoustic environments - the Mandarin system required us to address a number of new language-specific research challenges, such as tone modeling and the tokenization of (non-whitespace-separated) character
streams into lexical entries for language model creation. Constructing such a system, essentially from scratch, in a year was a major effort – managed almost single-handedly by research associate (now a first-year graduate student) Yan Huang – and we were delighted that the resulting evaluation system performed at highly competitive levels for mainland Chinese broadcasts. In the process, we gained significantly from the experience of applying our techniques to a non-European language.

The majority of our efforts in Rich Transcription have been devoted to the Metadata Extraction (MDE) task. The MDE effort has two main components: “Diarization”, which seeks to label acoustic sources in the audio stream (speakers, music, noise, ...) and which currently deals primarily with speaker segmentation and clustering, and “Structural Metadata”, which is concerned with the automatic detection of events such as sentence boundaries and disfluencies.

For Diarization, we continued work in improving the performance and increasing the efficiency of IDIAP’s speaker segmentation and clustering algorithm [1]. This included the re-engineering of our diarization software for greater efficiency and for cross-platform portability, and exercising the new codebase to explore system parameters and signal processing representations. Work focused on diarization of Broadcast News data, where the mix of different speakers, background music, commercials, and various recording environments presents a particularly interesting challenge to the system. It should also be noted that this same segmentation and clustering code forms the basis for the explorations of language-independence subword units referred to above; in that case we allowed the software to create much smaller segments, which should converge to consistent acoustic element classes rather than to speakers. This latter work represents a collaboration between the original author of the codebase, Jiendra Ajmera, who visited ICSI from IDIAP, another IDIAP visitor, Micha Hersch [9], and ICSI’s lead on this investigation, Chuck Wooters.

For Structural Metadata, the challenge is to detect “hidden” events implicit in the speech stream, which should aid in structuring the (otherwise) unformatted and unsegmented word stream that ASR systems produce. Currently the primary targets are segmentation of the speech into “sentence-like” units (abbreviated to SUs) and the detection (and potentially repair) of speech disfluencies. Our current approach to this problem has built on the fundamental work of Shriberg, Stolcke, and their colleagues at SRI on “Hidden Event” modeling [23], which incorporates a language model capturing sequential information about words and interspersed events, a prosodic model drawing on an inventory of prosodic features that cue events of interest, and a mechanism for combining these two sources, principally through an HMM framework.

Work this year on Structural Metadata has included:

- Language model explorations – We examined several different kinds of language models (basic N-grams, part-of-speech and other class-based models, as well as an explicit repetition model) alone and in various combinations, for improved metadata event detection. We also tried augmenting our limited supply of carefully annotated in-domain training data with data mapped from other sources (such as data extracted from the Web, or newsstream data where we inferred SU boundary information simply from punctuation).

- Machine-learning techniques to address the imbalanced data problem – A particular problem for the decision trees used in our prosodic modeling is that instances of the “non-event” greatly outnumber the events of interest (such as sentence boundary locations). We had traditionally used simple downsampling of the majority (“non-event”) class to
balance the classes, but found significant improvement from using bagging and ensemble approaches in MDE training.

- Expansion of our feature inventory for prosodic models – New features include the use of Fujisaki parameters and the existence of possible breath occurrences for phrase and sentence boundary detection in BN. We also examined speaker-related cues when modeling for CTS, and are beginning to explore other longer-context features.

- Alternative model combination techniques – We have examined a number of techniques for combining the various knowledge sources used in our MDE models. Techniques included the use of transformation-based learning, and of factored language models, which support the conditioning of N-grams on prosodic feature classes. We are also starting to explore maximum entropy approaches.

- Cross-corpus comparisons of component performance – We conducted a detailed comparison of the effectiveness of our various event detection models, exploring their relative contributions for CTS and for BN, as well as for the Meeting data described in later sections. Not surprisingly, the impact of the model constituents varied considerably across domains, reflecting differences in speaking style, well-formedness of utterances, and available training materials.

- Fragment detection – Because of their important role in signalling disfluent speech and the difficulties they present to ASR systems, we continued to explore automatic word fragment detection, using models incorporating new prosodic and voice quality features.

Unlike the core ASR work, which has a long legacy of government-sponsored research and associated infrastructure supporting it, Metadata Extraction is really an emerging area of research. Consequently, we also devoted considerable effort this year to infrastructure development, including working with support organizations such as the Linguistic Data Consortium and NIST on task definitions, annotation specifications, pilot annotation exercises, and testing of new scoring tools and metrics. While such work limited the time available to us for fundamental research, we believe that these efforts are necessary for constructing a solid foundation for EARS Metadata efforts, and one that can provide a coherent framework which should be of value to the greater research community.

The SRI+ICSI+UW team has taken a leadership role among the teams working on Metadata under EARS, both in planning and infrastructure and in fundamental research. Our degree of involvement was made clear in this year’s EARS Rich Transcription evaluations: we were the only site to participate in both the spring and fall evaluations of Diarization (one in conjunction with the spring speech transcription evaluation of BN and CTS, and the other as part of the fall metadata evaluations) and we were the only ones to participate in the full array of tasks – across all domains and all target types – in the fall.

In addition to the focus area projects described above, we have also embarked on several efforts that bridge the Rich Transcription research areas. One of the most interesting of these is an effort to use extracted Metadata to improve speech transcription by means of a feedback mechanism to the recognizer. As a first step, one of our Swiss visitors is working on using sentence segmentation and diarization information in BN to provide better chunks for re-recognition by the ASR engine. Another line of research, supported work primarily conducted by our partner site UW, involves the joint optimization of ASR and MDE processing. This work takes N-best
or lattice output from the ASR engine, determines the metadata (currently just SU boundaries) associated to each hypothesis, and then forms a "confusion network" incorporating the hypothesized words and interspersed metadata, all weighted by their associated probabilities. One can then determine the single most probable path through the joint ASR-MDE network.

5.1.2 Novel Approaches

As noted in last year's annual report, the objective of this project is to develop a new speech representation to replace the standard cepstral transformation of a local spectral envelope. This includes work on the acoustic "front end", and statistical modeling for the new features that are being generated. The purpose is to provide significant new improvements for speech-to-text transcription, with a particular emphasis on conversational telephone recognition. Given the infrastructure that was put in place last year, this year we were able to refine our methods and scale them up from a small vocabulary task to testing on large vocabulary conversational telephone speech. An important goal for 2004 will be to work with SRI to transfer this technology into a full system for participation in NIST's official Rich Transcription evaluation in the fall, a key milestone for this DARPA-funded project. We will of course continue to do more fundamental research on the basic methods, in collaboration with our research colleagues at IDIAP, Columbia, University of Washington, and SRI.

The core technical ideas of this effort have been to:

- Escape the dependence on the short-term (10-30 ms) spectral envelope - replace (or augment) it with alternative features, especially probabilistic ones (as opposed to simple transformations of local spectral energy).

- Use multiple front ends across time/frequency - in particular, representing local spectral information across longer stretches of time, and other more general functions of the time/frequency plane.

- Design optimal combination schemes for these alternative feature streams.

- Modify the statistical models to accommodate the new front ends. This may be particularly important for the features that are computed over temporal windows that greatly exceed the current practice.

The emphasis at ICSI has been on the first three of these, though this may shift later on in the project. Additionally, we have taken on the responsibility of integrating compatible ideas from our collaborating sites into a full system for conversational telephone speech, and then working with SRI to transfer the technology into a system that has many more performance-enhancing features.

It is often difficult to bridge the gap between small and large speech recognition tasks. The former are useful for quick prototyping and for directing research through diagnostics on system behavior. However, new techniques that have been shown to work on a small task may or may not improve performance for large-scale recognition. Nonetheless, it is also true that most techniques that have become "classic" methods for large-scale recognition were originally developed on much smaller tasks.
Our approach for this problem was to develop an intermediate task based on the recognition of the 500 most frequently used words in the Switchboard corpus. Specifically, we used utterances that contained at most 10% Out-Of-Vocabulary words, a restriction that still permitted us to use about half of the database. Given the high frequency of these vocabulary words, we were able to conduct many experiments with about 20 hours of training data; the larger vocabulary tests typically required significantly more in order to get word error rates low enough to be good predictors of performance for full systems that are trained on hundreds of hours of speech (up to 2000 hours for the 2004 fall evaluations).

Given this intermediate task, we progressed in 2003 from recognition of natural numbers, through the 500-words task, and finally to the full vocabulary conversational speech task. Each stage required significant experimentation, but the broader search through design possibilities was conducted with the simpler tasks. The overall project may then be likened to a pipeline, in which a range of exploratory methods developed for small vocabulary tasks are applied to a midsize vocabulary conversational task, and in which the best approaches are then applied to full vocabulary CTS.

The feature set that we worked on in 2003 was a variant of the Temporal Pattern (TRAP) features [8] called Hidden Activation TRAPS (HATS) [3]. In this approach, hidden layers trained in 3-layer multi-layer perceptrons (MLP) for phone discrimination (using temporal energy patterns over 500-100 ms in 1 or more critical bands) are used as nonlinear feature extractors. These features are then used as input to a combining network, the output of which is then gaussianized and orthogonalized for use as observations for hidden Markov models. As is usually the case for the MLP training methods we use, the outputs of the combining network may be viewed as posterior probabilities of the classes being discriminated. As such they can also be combined with the outputs from a feature transformation network that uses conventional features as input (e.g., PLP), using one of several plausible methods (e.g., arithmetic or geometric averaging) prior to the later steps. In practice, we have often found a weighted linear combination to be useful, where the weights are proportional to the inverse of the entropy calculated from the vector of posteriors. Finally, the transformed final posterior estimates are typically concatenated with baseline features (e.g., PLP and derivatives, or a linear transformation of these features) to form an extended feature vector for use in training and recognition [19, 20]. Figure 1 shows this process.

Early in 2003, we demonstrated that these features significantly improved performance for the Numbers task developed by OGI, which we have used for many experiments over the years. By mid-year, we were able to derive similar advantage from the conversational telephone speech (CTS) 500-word task. In both cases the training data were primarily CTS. This was done so that the detailed design choices made during the feature development would be more likely to generalize beyond the Numbers task.

These experiments were done using roughly 20 hours of speech. In order to progress to the full vocabulary task, we had to expand our training (both for feature-generating MLPs and for HMMs) significantly. In fact, this expansion represents what is most likely the most significant challenge, both in terms of computational limitations and in terms of preserving the advantage that we are seeing when using our new features in the smaller tasks. The latter point is an issue because modern state-of-the-art laboratory speech recognition systems have many features that permit them to take advantage of large amounts of training data (for instance, speaker-adaptive training and multiple passes of MMIE training).

With the next larger training set (64 hours) and using the extended feature vector described
above, concatenating regular PLP feature with TRAPS-based features, we saw an improvement in word error rate from 43.8% for the PLP-only system to 40.6% for the one with combined features, a relative 7.3% error reduction. In a second set of experiments at ICSI, incorporating further work on both the MLPs (for instance, using the HATS discussed above) and the system tuning, as well as the commonly-used adaptation called MLLR, we observed a significant improvement (8.9% relative) in word error rate. This was a reduction from a PLP baseline of 35.8% to 32.6%, for the the NIST-standard Eval 2001 test set given 64 hours of training.

By the end of 2003, we had progressed to larger and larger systems incorporating more characteristics of the full SRI system, and also training both HMMs and MLPs on larger amounts of data. The best such system to date incorporated MMIE training, 4-gram and duration rescoring, and was trained from 420 hours of speech for the HMMs and 260 hours for the MLPs. In this case, the new features reduced the error rate from 25.6% to 23.5%, an 8.3% relative reduction, again for the Eval 2001 test set. Reaching this point took months of experimentation to achieve, but ultimately it appears that the new features can indeed scale to larger training sets and better recognition systems. This is an extremely promising result that gives us confidence that we should be able to have a significant effect on the NIST evaluations in the fall. Given this result, SRI has committed to working with us to incorporate these features in their system, and other sites (researchers at IBM, BBN, and Cambridge) have expressed similar interest. It is worthwhile noting that these kinds of error reductions are viewed as major in this community, where huge amounts of effort are expended to achieve a few percent error reduction for difficult tasks like CTS recognition.

Much of the effort described above is necessarily incremental, since it is quite difficult to extend novel approaches from feasibility demonstrations on small tasks to embedding in complete large systems. However, we are always also working on the next "wave" as well. If discriminant features trained from critical band temporal trajectories show promise, what about training such features from more general functions of the time-frequency plane? We have been employing a framework based on a set of two-dimensional Gabor filters with varying extents in time and
frequency and varying ripple rates to analyze a spectrogram [15, 14]. These filters have some characteristics in common with the responses of neurons in the auditory cortex of primates, and can also be seen as two-dimensional frequency analyzers. Promising results have been obtained in a noisy digit recognition task, especially when this analysis method was combined with more conventional analysis. Work is ongoing in the use of this approach for larger-vocabulary recognition tasks, and in the use of the Gabor filters in a multi-stream architecture in which multiple classifiers, each using different representations, operate in parallel. An important part of this work is the use of automated feature selection to find appropriate Gabor filters.

5.2 Speech from Meetings

The ICSI Speech group continues to invest considerable effort in the analysis of speech from meetings. Work in this area ranges from corpus creation and annotation, to core signal processing and speech recognition, to analysis of higher-level structure and content. This year’s efforts were supported by a continuing NSF-ITR in “Mapping Meetings” as well as our participation in two European initiatives: the EU framework 5 program M4 (Multi-Modal Meeting Manager), and the Swiss program IM2 (Interactive Multimodal Information Management), funded by the Swiss NSF and managed by IDIAP. In addition, beginning in January 2004 our meetings research is bolstered by a new award as part of the 15-site consortium AMI (Augmented Multi-party Interaction), an EU framework 6 program.

The sections below describe some of the highlights of this year’s Meetings effort.

5.2.1 Corpus Development

In last year’s annual report we described our collection of a corpus of data from natural meetings that occurred at ICSI over the period of 2000-2002. The corpus contains audio recorded simultaneously from head-worn and table-top microphones, word-level transcripts of meetings, and various metadata on participants, meetings, and hardware. Such a corpus supports work in automatic speech recognition, noise robustness, dialog modeling, prosody, rich transcription, information retrieval, and more. Transcription and quality control stages were completed last year, and the annotated corpus was delivered to the Linguistic Data Consortium (LDC) in mid-2003; it is now being offered to the research community through the normal LDC licensing procedure. Further details on the corpus can be found on our web site: http://www.icsi.berkeley.edu/Speech/mr. Although the corpus has now moved out of our hands, work continues on improving the data provided – both incorporating minor corrections and processing a handful of additional meetings not included in the official release – and in enriching the corpus through the addition of extra levels of annotation (such as the Dialog Act labels described below) and of supporting tools. For example, in order to support transcription of meetings data, we had earlier modified the publicly available Transcriber tool, creating a new version (“channeltrans”) which facilitates multi-channel processing. Channeltrans is currently being modified to directly read and write data in ICSI’s MRT format, an XML-based format designed expressly for our meeting data and the format used for the transcripts in our released corpus.

5.2.2 Dialog Acts

The goal of this work is to automatically label dialog acts in order to provide better input to systems designed to characterize or summarize meetings. For example, dialog acts may be used
to spot locations of agreement/disagreement, floor-grabbing, topic shift, etc., and may be used to refine language models both for act-specific word usage and to model turn-taking patterns over the course of a meeting.

Here we define a dialog act as the characterization of the function or role of an utterance in the context of the conversation. A set of 58 tags was defined for this work, based on the Switchboard-DAMSL conventions [12] and refined over time by ICSI’s annotation team to reflect phenomena observed in the meeting data. The basic utterance types of statement, question, and backchannel (such as “uh-huh”) form the primary layer of description, with additional tags providing multiple levels of refinement.

We have completed Dialog Act and Adjacency Pair labels for the 75 officially released meetings from the ICSI Meeting Corpus. In addition to the labeled corpus, we provided:

- word time-alignments from forced alignment (using SRI’s speech recognizer)
- error codes indicating isolated anomalies
- miscellaneous comments from labelers
- a set of classmaps for mapping the large number of observed label sequences to meaningful smaller numbers of classes
- statistics on interlabeler reliability
- a detailed manual describing the labeling
- other documentation on issues in using the corpus

This DA corpus is now being shared with research partners at other sites, and plans are being made for a more general public release of the material.

5.2.3 Detection of Meeting Hotspots

Recent interest in the automatic processing of meetings is motivated by a desire to summarize, browse, and retrieve important information from lengthy archives of spoken data. One of the most useful capabilities such a technology could provide is a way for users to locate “hot spots”, or regions in which participants are highly involved in the discussion (e.g., heated arguments, points of excitement, and so on). Such regions are likely to contain important information for users who are browsing a meeting or for applications of information retrieval. We researched the following two questions:

- Can human listeners agree on utterance-level judgments of speaker involvement?
- Do judgments of involvement correlate with automatically extractable prosodic cues?

To address the first question we conducted a study in which human subjects were asked to rate utterances with respect to involvement. We found that despite the subjective nature of the task, raters showed significant agreement in distinguishing involved from non-involved utterances. We also found a difference in ratings depending on whether raters were native or nonnative speakers of the language – which may reflect language differences, cultural differences, or both.
To address the second question, we correlated acoustic features based on F0 and energy with the human ratings of involvement. These acoustic features were extracted and (where applicable) normalized completely automatically based on previous work at ICSI (funded by the DARPA Communicator project). We found remarkably reliable acoustic cues to involvement, based on F0 and energy values. Furthermore, it is likely that this is a general effect over all speakers (rather than a correlation between speaker prosodic values and speaker tendency for involvement), because we found that the most affected features of an individual speaker were similar to the most affected features that were computed over all speakers.

Taken together, these results suggest that hotspots, as defined via involvement level of utterances, can be fairly reliably identified by humans, and thus could be an important construct to label in our meeting maps. The automatic detection of these regions can take advantage of prosodic cues, as we have found in this study and are continuing to pursue. In future work, we are planning to examine lexical cues (using both true and automatically recognized words), looking at hotspot identification directly (rather than via involvement at the utterance level), investigating subclasses of involvement (e.g., arguments versus jokes), and further exploring acoustic cues. By increasing our database of hotspots, we will also be able to conduct machine learning experiments to predict involvement on unseen data.

5.2.4 Discourse markers of meeting content and social structure

Our collaborators at University of Washington and Columbia University are finding ways of identifying topics and topic shifts. Our work is intended to identify points of maximal information within topics and speaker-perceived relations between one topic and the next.

Particularly relevant is the word “so”. In the literature this discourse marker has been identified with introducing main points and conclusions and with relinquishing the floor.

The literature contains successful attempts at identifying prosodic profiles for specific uses of other discourse markers. It is our goal to derive similar profiles for discriminating the different meeting-structuring uses of “so” and to distinguish these from non-discourse marker uses (such as “and so forth”).

Our findings in this area were:

1. Agenda setters tend to use more topic-structuring and topic-concluding “so” markers than do other meeting participants.

2. The frequency of topic-structuring markers will differ for an individual as a function of speaker role (that is, more of them in meetings in which that speaker is the agenda-setter than in meetings when he/she is not).

3. We replicated the finding of turn-final “so” being used to relinquish a turn, but noticed that unlike prior work, these uses of “so” sometimes occur as much as one second into the new speaker’s turn (suggesting they may sometimes not be cues to turn change, but merely correlates of them).

4. We discovered a new use of “so” in turn-taking, which is to claim the floor. This use is very common in our meetings but not reported in the conversation analysis literature.

Our findings in items 3 and 4 suggest a larger point, which is that some properties of meetings (e.g., more competition for the floor, more time constraints for relevance, and more limited
bandwidth for monitoring every other speaker's nonverbal cues) may not be covered by existing theories of discourse or conversation analysis. That is, our findings on meetings may point to needed extensions of well respected theories developed for more usual types of conversation.

5.3 Spoken Language Systems

In much of what we do in our speech research, we measure performance with standard test sets (e.g., NIST evaluations of CTS and BN speech recognition in EARS). However, we have learned a lot by applying new technological approaches to complete systems and getting experience with more practical aspects, even though we are not concerned with building products. In particular, it gives us the ability to see how individual components that we have developed work in concert with other parts of a more complete system. The Berkeley Restaurant Project (BeRP) dialog system [13] gave us this kind of experience. More recently, as reported in previous annual reports, we have been working with DFKI and other German labs to develop what is called SmartKom (see http://www.smartkom.com). This project, funded by the German Ministry of Education and Research, has had the goal of developing a multi-modal dialog system that can assist humans by interacting with speech, gesture, and facial expression. Aside from specific research explorations such as noise and acoustics compensation, robust speech recognition features, and semantically rich language analysis, the focus of the ICSI work was to provide an English-language version of the system.

The development of an English-language SmartKom Mobile system verified the language portability of the SmartKom architecture and facilitated the demonstration of SmartKom at international conferences. Staff and visiting researchers at ICSI and staff at DFKI took the lead roles in the creation of the English language system, and important contributions came from several other SmartKom partner sites. The English speech recognizer was developed completely independently from the German one; it is a hybrid connectionist system descended from the one used in BeRP. All other modules in the English SmartKom Mobile system were based on, or identical to, modules in the German-language SmartKom Mobile. The modular architecture of SmartKom greatly eased porting to English by encapsulating language dependencies in specific modules. Most SmartKom modules required no modification to support English. Of the modules requiring modification, only speech recognition and speech synthesis required significant changes to software source code. The speech analyzer (which parses the recognized speech) and text generator (which creates the system’s output sentences) required only a change in their template (grammar) files and otherwise used the same software engines for both German and English. The lexicon module had to be aware of the current language in order to provide the correct word pronunciations. Some displayed text provided by the pedestrian and vehicle navigation modules (such as tourist site information and map labels) was translated to English. The speech analyzer outputs a language-independent semantic representation of the user input, and so modules which tracked dialogue state and user intention did not need to be language-aware.

The English system was demonstrated at several major conferences dealing with speech processing or human-computer interaction (EUROSPEECH 2001, ICSLP 2002, ICMI-PUI 2003) and during an ICSI Open House that was part of the UC Berkeley BEARS conference, 2004.
5.4 Speaker Recognition: Modeling Idiosyncrasies in Speaking Behavior

This year, ICSI embarked on a new project in speaker recognition, seeking to characterize the idiosyncratic nature of an individual’s speech. Speaker-characteristic information is encoded in multiple levels in the speech signal: the low-level acoustics of spectral patterns, prosodic features, pronunciation choices, word usage, and idiosyncratic vocal gestures. Yet most speaker recognition systems today use only the lowest level acoustics, typically relying on standard signal-processing techniques that break the speech stream into a series of 10-20 millisecond “frames” modeled as essentially independent events. Recently, based on a series of strikingly successful pilot experiments, there has been a surge of interest in the exploration and incorporation of speaker information sources “beyond the frame”. This project explores higher-level sources of speaker information. The goal is two-fold: to conduct fundamental exploratory research to discover speaker-distinctive features and encode them into richer, more informative speaker models; and to evaluate the utility of these feature sets and models for speaker recognition and other speech technology applications.

The feature discovery effort proceeds in parallel along two complementary tracks: a “knowledge-based” track, building on existing linguistic constructs and guided by insights from psycholinguistics and human performance studies, and a more speculative “data-driven” approach, seeking idiosyncratic “vocal performances” – spectro-temporal patterns with high speaker-characterizing power, independent of linguistic constraints. This two-pronged approach pairs a low-risk, expert-guided search with a higher-risk, but potentially higher pay-off, exploration. This provides a natural contrast for assessing progress and an opportunity to combine two very different information types into a more effective whole.

On the theoretical side, the new effort should lead to a better understanding of what makes a speaker’s voice and speech behaviors unique, what constitutes normal within-speaker variation, and what dimensions are important for modeling speaking style. Further, the resulting speaker models should have a number of practical applications, not only for improving speaker recognition systems, but for numerous other speech technologies that could benefit from speaker-sensitized processing: providing more speaker-focused models for automatic transcription, facilitating analysis of discourse structure and speaker behavior, and supporting more personalized human-computer interaction.

While the project has only begun recently, we have already completed some relevant experiments. For example, in contrast to the typical “bag of frames” models, which treat every speech frame as an essentially independent event, ignoring temporal dependencies between frames and instead creating a single large Gaussian mixture model of generic speech, we have examined the performance of hidden Markov models (HMMs), which allow for more incorporation of the temporal aspects of speech. HMMs have standardly been used in text-dependent, password-based speaker verification systems, but are less common for the sort of text-independent systems of interest here, where speaker recognition must use whatever unconstrained conversational speech is available to the system. By constructing word-specific HMMs for a small number of words commonly encountered in conversational speech (backchannels, discourse markers, hesitation words), we were able to demonstrate strong speaker characterizing power even though only a tiny fraction of the words in the conversation were being used. Furthermore, by coupling the relative frequencies of these words with the quality of the acoustic match as modeled by the HMM, performance improved even more: clearly both a speaker’s choice of backchannel (“uh-huh”, “yeah”, “right”, “really”, ...) and what he sounds like when he says one provide
important clues to speaker identity. Preliminary results such as these provide further evidence of the benefits of examining higher level phenomena for the purpose of speaker recognition.

5.5 Computer architectures for speech processing

Automatic speech recognition provides a natural interface to small form-factor computers (such as PDAs) since keyboards and large displays are absent on these platforms. However, large vocabulary, robust ASR requires hardware resources far beyond those available on current PDAs. Emerging architectures, such as Vector IRAM at UC Berkeley, and Imagine at Stanford, provide a partial solution by delivering very high performance for relatively little expenditure of power. However, for speech recognition to take advantage of these architectures, the components of the system must be redesigned with the new systems in mind.

We are currently adapting the workstation-based ASR system used at ICSI to run efficiently on these architectures. Two out of the three major components of ICSI’s speech system, the acoustic front-end and the phoneme probability estimator, contain computational kernels that are very regular (FFT and matrix-matrix multiply, respectively). These components run extremely efficiently on both architectures. The third component, the decoder, consists of a highly pruned (and therefore irregular) search through all possible utterances. Thus, our primary focus is on this portion of the speech system.

With a small vocabulary system, it is not necessary to share states among words; rather, one can evaluate all the words separately. On vector architectures, we can arrange the words into batches with total length equal to the vector length. This allows an efficient, regular implementation on vector architectures.

Large vocabularies are more problematic, since it is inefficient to evaluate all words in the vocabulary. Instead, the algorithm should only evaluate words that have a high probability of being matched. Also, the algorithm should reuse the results of previous words when computing the current word (e.g., if the word “architecture” has already been processed, most of the work can be reused for the word “architectural”). Several strategies are currently being investigated for large vocabulary decoders, including maximally separated search orders and randomized search orders.

5.6 Physiologically motivated speech processing

Much of our effort on speech signal processing has always been motivated by psychoacoustic results; that is, based on measurements of phenomena of human hearing. However, we have generally not made much use of direct models of auditory physiology, since we assumed that whatever was relevant from these measurements would be observed in the behavioral results. However, there is interesting work going on in a number of areas that derives more direct inspiration from physiological (and in particular, neural) mechanisms. In an earlier section of this speech report, we mentioned our use of Gabor functions as a simple model for observations from the primary auditory cortex of small mammals; we hope to work more extensively in this area in the future. Additionally, in collaboration with colleagues at Infineon in Germany, we have been working to integrate a detailed model of the human cochlea into one of our speech recognition systems. Experiments have been performed using this model, and future plans are to model the cochlear nucleus and inferior colliculus as well.
5.7 ICT4B

ICT4B (Information and Communication Technology for Billions), is a UCB-led project that aims to develop key technologies and infrastructure to enhance projects that have the goal to bring information and communication technology to developing regions. In particular, it seeks to address key challenges in cost, deployment, and power. Additionally, its goal is to provide support for populations with low levels of literacy and multiple languages, which makes the development of an appropriate speech interface crucial. The speech component of ICT4B, based at ICSI through a subaward, seeks to develop the speech toolkit required to drive ICT4B applications.

As part of this project, we are now working on a data collection for Tamil. A small data collection application will run on a laptop, and will be used to collect a speech from native Tamil speakers on campus. This will be used to bring up an initial recognizer.

References


6 Berkeley Center for the Information Society

The Berkeley Center for the Information Society is a research center started at the International Computer Science Institute to bring computer science and social sciences together. This interdisciplinary connection will be very critical for the next stage of the technological development.

The center is directed by Dr. Pekka Himanen and its research board is chaired by Prof. Manuel Castells. Key areas include: (1) challenges of the global information society and different models of responding to it; (2) the use of the open-source model for social projects; and (3) enhancing equal social opportunities with IT. Among the current projects in the above areas are (1) the comparison of the Silicon Valley, Finnish, and Singapore information society models by Pekka Himanen, Manuel Castells, AnnaLee Saxenian and their group; (2) research on the application of the open-source model to social movements by Jerry Feldman, Steve Weber and others; (3) the digital opportunities program called “Berkeley Foundation for Opportunities in Information Technology”, led by Orpheus Crutchfield; and (4) A pilot project on IT in Ghana, led by G. Pascal Zachary.

The “Berkeley Foundation for Opportunities in Information Technology” (BFOIT) program continues to play a lead role in providing IT career help for underserved groups in the Bay Area. One indication of its acknowledged excellence is a multi-year grant from the Elizabeth and Steven Bechtel, Jr. Foundation, to expand our year-round and summer programs and serve more students. BFOIT has continued the year-round academic enrichment program with college coach Barbara Austin. We are encouraged that we have had a steady group of 15-20 all year. Several of our students have been accepted at engineering schools, and are interviewing for full scholarships. BFOIT has begun to put together a tangible pipeline, starting from middle school straight through college. In the summer, we will begin a pilot program called SCI-FY, (Science and Computer Camp for Youth), which will effectively double the number of students BFOIT works with. The camp will be taught by long-time BFOIT instructor Guy Haas, former Sun Engineer, and tutor Ayodele Harrison. We are working in conjunction with Supervisor Keith Carson’s office, serving Berkeley and Oakland students. BFOIT is partnering with a premiere mentoring program called Sage Fellows, a former UC program. We hope that this partnership will yield real mentors for our students as they strive for opportunities in computer science and engineering.

The Second Annual “New Technologies Day” at Microsoft was held this Spring. It grew from 55 to 68 participants and one of the vice-presidents asked us to keep expanding the program. Several parents chaperoned the all-day event and they and the students were very pleased with the event which involves students in North Carolina, Texas and Silicon Valley. There are talks emerging around starting another BFOIT unit in the San Jose area, initiated by an active parent in the current program. We are working out the logistics for a Summer 2004 start. We are waiting to hear the results of a special grant to an individual BFOIT student to receive honed, personalized college coaching for admissions and scholarships; prospects look good!

The pilot project in Ghana has, at least temporarily, been suspended due to a lack of continued funding. In addition to some individual successes, the most lasting result was the establishment of a new machine assisted curriculum at Ashesi University. This was done in collaboration with Roger Schank of CMU, who developed the course material and with Patrick Awuah, the president of Ashesi University.

In the fall 2002, the fellows and visiting fellows of the Center included Prof. Manuel
Castells, Orpheus Crutchfield, Prof. Peter Evans, Prof. Prof. Jerry Feldman, Dr. Antti Hautamaki, Dr. Pekka Himanen, Prof. Youtien Hsing, Prof. AnnaLee Saxenian and Prof. Harley Shaiken.

The Center has also formed a group of doctoral students in this research area. Two UCB graduate students are working full time for BCIS - Joyojeet Pal from SIMS and David Thaw from Political Science. In the summer of 2003, BCIS supported six students working on IT and Society. In the fall, these students and others had a working seminar, designed to build a community of UCB grad students studying and carrying out IT activities to benefit society, particularly underserved people and areas.

In 2003, BCIS undertook a study of Civil Communities of Practice for Sitra, the Finnish development agency. The resulting report is available and is forming the basis for specific efforts in Finland, the Central Valley and elsewhere.

The center's research board includes top people from different backgrounds, such as Linus Torvalds (the creator of the Linux operating system) and Barbara Simons, recently President of the ACM. The goal of the Berkeley Center for the Information Society is to develop an interesting network of interaction between the top academic, business and civil society people, who share an interest in the social good of the information age.

Near the end of 2003, Prof. Anna Saxenian of BCIS was appointed Dean of the UCB School of Information Management and Systems. This, plus the continuing cooperation with Citris and the ICT4B project ensures strong campus participation in BCIS activities.

References

7 Robust Video Compression based on Distributed Source Coding techniques

Abhik Majumdar and Prof. Kannan Ramchandran worked in 2003 on a video coding project at ICSI. This project builds on a novel video coding paradigm called PRISM [2] which is based on distributed source coding principles from multi-user information theory. Today's state-of-the-art video coders (such as MPEG, H.26x, Motion-JPEG, etc.) can provide either high compression efficiency (MPEG, H.26x) or low encoder complexity and robustness (Motion-JPEG). The design goals of PRISM are to provide all of these desirable features simultaneously. This is accomplished by (i) enabling the complex motion-estimation task to be shifted from the encoder to a network device; and (ii) leveraging channel coding concepts that underly distributed compression techniques. These features make the PRISM codec ideally suited to the “uplink” video transmission applications, such as video streaming from cellphones, PDAs, etc.

The highlights of the last twelve months of this project include:

- Improved compression efficiency of the PRISM codec: we have improved the “pure” compression efficiency by the order of 1 dB in Peak SNR for typical test sequences. This is accomplished through the use of improved codes in PRISM (see next item also).

- Significantly more flexibility in terms of the range of encoding rates (the original version was limited to 1 bit per sample type of operation). This is accomplished by generalizing the original trellis-based codes to a much richer class of generalized coset codes (termed “Multilevel Coset Codes”).

- Improved resilience to drift in conventional predictive video coders (such as MPEG, H.26x etc) through the use of a side-channel. This is accomplished by using a PRISM-like bit-stream on the side-channel to enable refinement of the video sequence decoded by the conventional coder.

7.1 Summary of Results using Multilevel Coset Codes for PRISM

For performance tests we compared three video coders: Multilevel and Trellis codes-based PRISM and H.263+. Preliminary simulation results indicate that the Multilevel Codes based PRISM codec outperforms the original trellis coded implementation by about 1 dB at high rates and is only 1.4 dB worse than H.263+. Apart from pure compression performance, we also tested the robustness characteristics of the three video coders using a wireless channel simulator obtained from Qualcomm, Inc. For loss rates between 2-16%, we found that both implementations of PRISM clearly outperform H.263+, with the Multilevel-codes based PRISM about 4-7 dB better than H.263+ for the loss rates shown. We also found that the Multilevel Codes based PRISM performs consistently better than the trellis coded implementation at the loss rates investigated. Figure 5 shows the snapshot of the 15th (last) frame of the Football video sequence decoded by the H.263+ and (Multilevel) PRISM coders. For further details see [1].

---

2 PRISM stands for Power efficient Robust high compression Syndrome-based Multimedia coding.
3 Free Version of H.263+ obtained from Univ. of British Columbia.
7.2 Improved Resilience to Drift through the use of a Side-Channel

In this project we address the problem of drift in conventional predictive video coders (such as MPEG, H.26x etc) through the use of a PRISM-like bit-stream on a side-channel to enable refinement of the video sequence decoded by the conventional coder. The system configuration is shown in Figure 6. The MPEG encoder and decoder work independently of the side-stream encoder and decoder. Thus, if the client has access to only the MPEG decoder then it will be able to decode the MPEG stream, but if it has access to both the MPEG and side-stream decoders then it will be able to decode to a better quality. As shown in Figure 6, the side-stream encoder encodes the current block (X) assuming that the side-stream decoder has access to the MPEG decoder’s reconstruction of the current block (\(\tilde{X}\)). The encoder estimates the correlation between X and \(\tilde{X}\) through a correlation estimation algorithm. We have developed the correlation estimation algorithm using techniques illustrated in [4]. The side-stream encoder encodes \(\tilde{X}\), treating \(\tilde{X}\) as the side-information present at the decoder, by exploiting Wyner-Ziv [3] coding principles. When decoding succeeds, the side-channel decoder decodes to \(\tilde{X}\) (the intra value of X), thus stopping the drift. We assume an independent packet erasure model for both the channel for the MPEG stream and the side-channel.

References

