INTERNATIONAL COMPUTER SCIENCE INSTITUTE

ACTIVITY REPORT 2002

INTERNATIONAL COMPUTER SCIENCE INSTITUTE
1947 CENTER STREET, SUITE 600
BERKELEY, CA 94704-1198 USA

PHONE (510)-666-2900
FAX (510)-666-2956
E-MAIL info@icsi.berkeley.edu
http://www.icsi.berkeley.edu

PRINCIPAL 2002 SPONSORS

Intel
Nortel
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 2002 SPONSORS

European Media Lab
European Union (via Sheffield University)
IBM
Infineon

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. Elwyn Berlekamp (Chair), UC Berkeley
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, Comventures
Prof. Richard Karp, ICSI and UC Berkeley
Mr. Pedro Lizcano, Telefonica
Prof. Jitendra Malik, CS Division Associate Chair, UC Berkeley
Prof. Nelson Morgan (Director)
Dr. Ilpo Reitmaa, TEKES
Prof. Wolfgang Wahlster, DFKI GmbH

SENIOR TRUSTEES

Prof. Jerome Feldman, ICSI and UC Berkeley
Prof. Gerhard Goos, U. Karlsruhe
Prof. David Hodges, UC Berkeley, past ICSI Chair
Dr. Ron Kay, former acting Director of ICSI (until 1988)
Prof. Heinz Schwärtzel, T. U. Munich
Prof. Norbert Szyperski, U. Cologne; former head of the GMD; Founder of ICSI.

2002 VISITORS IN SPONSORED INTERNATIONAL PROGRAMS

NAME	COUNTRY	GROUP	AFFILIATION
Mikko Harju	Finland	Speech	TEKES
Antti Hautamaki	Finland	BCIS	SITRA
Lauri Hetemaki	Finland	BCIS	TEKES
Pasi Koistinen	Finland	Haas	TEKES
Jan-Ola Ostman	Finland	FrameNet	TEKES
Jussi Leppanen	Finland	Speech	TEKES
Kira Lopperi	Finland	Haas	TEKES
Kimmo Parssinen	Finland	Speech	TEKES
Panu Somervuo	Finland	Speech	TEKES
Seppo Raudaskoski	Finland	AI	TEKES
Katri Seppala	Finland	AI	TEKES
Krista Varantola	Finland	AI	TEKES
Ernst Althaus	Germany	Algorithms	DAAD
Sven Behnke	Germany	Speech	DAAD
Hans Boas	Germany	AI	DAAD
Hans Cycon	Germany	Speech	FV
Klaus Dolzer	Germany	Speech	DAAD
Christoph Erlen	Germany	Haas	CDTM
Christian Esser	Germany	Haas	CDTM
Stefan Harrer	Germany	Haas	CDTM
Werner Hemmert	Germany	Speech	Infineon
Phillipp Jostarndt	Germany	Haas	CDTM
Christoph Karg	Germany	Networks	FGAN
Michael Kleinschmidt	Germany	Speech	U. Oldenburg
Birger Kollmeier	Germany	Speech	U. Oldenburg
Marek Musial	Germany	AI	DAAD
Thilo Pfau	Germany	Speech	DAAD
Volker Roth	Germany	Networks	DAAD
Petra Steiner	Germany	AI	DAAD
Esther Teo	Germany	Haas	CDTM
Nikolaus Von Taysen	Germany	Haas	CDTM
Klaus Wehrle	Germany	Networks	DAAD
Mattias Westermann	Germany	Algorithms	DAAD
Britta Wrede	Germany	Speech	DAAD
Victoria Abreu	Spain	Speech	MCYT
Laura Docio	Spain	Speech	MCYT
José Gonzalez	Spain	Networks	MCYT
Marc Gratacos	Spain	Speech	MCYT
Cristina Lamana	Spain	BCIS	MCYT
Carlos Subirats	Spain	AI	MCYT
Micha Hersch	Switzerland	Speech	IM2

CDTM: Center for Digital Technology and Management
DAAD: Deutscher Akademischer Austauschdienst
EML: European Media Lab
TEKES: Finnish National Technology Agency
FGAN: Forschungsgesellschaft für Angewandte Naturwissenschaften e. V., http://www.fgan.de/FGAN/En/
MCYT: Ministerio de Ciencia y Tecnologia, Estado de Politica Cientifica y Tecnologia
IM2: Interactive Multimodal Information Management, National Centre of Competence in Research, Switzerland

INSTITUTE OVERVIEW – FEBRUARY 2003

The International Computer Science Institute (ICSI) is an independent, nonprofit basic research institute affiliated with the University of California 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 four major research areas: Internet research, including Internet architecture, related theoretical questions, and network services and applications; 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; and natural speech processing. Additionally, as of September 2002 we have a new activity focused on 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.

INSTITUTE SPONSORSHIP – 2002 FISCAL YEAR (SAME AS CALENDAR YEAR)

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.

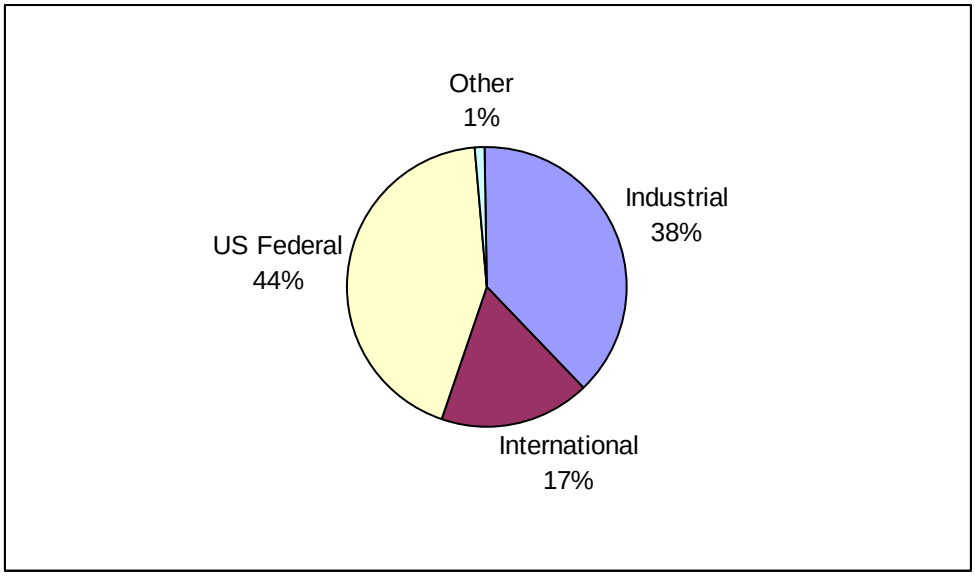

Figure 1: Distribution of sources of ICSI revenue for 2002.

US Federal funding comes from a range of grants that support research Institute-wide. Most of this funding comes from the National Science Foundation and DARPA. International support in 2002 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 came from the European Media Lab and the European Union. Industrial support in 2002 was primarily provided by Intel, Nortel, and Qualcomm, with additional sponsorship from Infineon and IBM. We also benefited significantly in 2002 from prior funding from AT&T.

Revenues increased in 2002 (about 14% over 2001), to roughly $8.1M for the year.

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.

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.

Most 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, and administrative support functions for ongoing operations and special events. Aside from these groups, a Contracts and Development officer handles proposal preparation and submission and external contracts.

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, and the Speech Group. Consistent with this organization, the bulk of this report is organized along these lines, with one sub-report for each of the four groups. A new activity that began in 2002 was the Berkeley Center for the Information Society (BCIS). This group explores the impact of IT on society, and is initially based on collaboration among Prof. Feldman, Prof. Castells of the UCB Sociology Department., and Dr. Pekka Himanen of Finland.

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, 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.

SENIOR RESEARCH STAFF

The previous section 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, 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 2003) 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.

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, such as word sense disambiguation, computer-assisted translation, and information extraction.

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.

MARK HANDLEY (NETWORKING): scalable multimedia conferencing systems, reliable multicast protocols, multicast routing and address allocation, and network simulation and visualization.

HYNEK HERMANSKY (SPEECH): acoustic processing for automatic speech and speaker recognition, improvement of quality of corrupted speech, human speech communication**.** (Also with the Oregon Graduate Institute).

RICHARD KARP (ALGORITHMS AND NETWORKING): mathematics of computer networking, computational molecular biology, computational complexity, combinatorial optimization**.**

NELSON MORGAN (SPEECH): signal processing and pattern recognition, particularly for speech and biomedical classification tasks.

SRINI NARANAYAN (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, particularly 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.

RESEARCH GROUP REPORTS

2002 was another difficult year for the world of technology, as we continued the global slowdown. While we certainly have been affected, ICSI has not only weathered these changes but in fact has expanded in some areas. The decline of industrial funding was largely compensated for by success in obtaining new competitive Federal grants. Significantly, the Networking group strongly shifted over towards Federal funding, and the speech activities expanded mainly due to a new DARPA program. In this report, we have described our research in terms of the four principal research groups in place during 2001: Networking, Speech, AI, and Algorithms. We will also describe the new Center (BCIS) as well as a few other activities that did not fit neatly into these categories.

RESEARCH GROUP HIGHLIGHTS

The following are a selection of key achievements in our research groups for the year 2002, 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.

NETWORKING

- Internet Security, National Studies: CSTB/NRC report on "The Internet Under Crisis Conditions: Learning from September 11"; DARPA report "Large Scale Malicious Code: A Research Agenda". Also, paper on "How to Own the Internet in Your Spare Time" that appeared in the USENIX Security Symposium has had 3.2 million hits and 86,000 downloads.

- Towards New IETF Standards in Congestion Control for Transport Protocols: TFRC, a congestion-control method applicable for streaming media and other non-bulk transfer applications (Proposed); DCCP, a congestion-control method for UDP, in process.

- Distributed systems: (With Algorithms) The large NSF ITR (five-year, 12M grant from the NSF for the Infrastructure for Resilient Internet Systems (IRIS) project. Based on ICSI-developed distributed-hash-table technology.

- XORP: The eXtensible Open Routing Platform project made its first public release of its code in December.

- Sally Floyd and Mark Handley now on Internet Architecture Board (IAB).

ALGORITHMS

- Computational genomics work: haplotyping, transcriptional regulation, motif finding, analysis of gene expression and gene regulation, and protein sequence comparison.

- Approximation algorithms: results on hardness of approximation and new approximation algorithms for a number of classic problems of network design.

- As noted previously, joint successful proposal process (with Networking) leading to IRIS project on peer-to-peer networks, with collaborators from UCB, MIT, Rice, and NYU. This was one of only 7 "Large" ITR projects funded this year by NSF.

- In 2002 the combinatorial games research group achieved an improved understanding of several kinds of loopy games.

AI

- FrameNet SW readied for international projects; 55 projects wish to use data (preliminary version released) for purposes including machine translation, information extraction (IE), frame semantic parsing, and teaching English as a second language. FrameNet frames and lexical units have been expanded with IE algorithms to produce good precision and recall with no further domain-specific effort.

- Significant advances in the use of Neural Grammar in courses and in applications, both linguistic and computational. Graduate course cross-listed in CS and linguistics.

- Extended cooperation with EML (Heidelberg, Germany), completed first semantic parser.

- Collaboration with UCB and Parma for behavioral and imaging studies, which produced evidence supporting Neural Theory of Language.

- Cooperative agreement with ARL to provide access to supercomputing for large-scale simulations of neural models of reasoning and memory.

SPEECH

- Large vocabulary speech recognition: Two DARPA "EARS" awards - The first was one of only three awarded in the "Rich Transcription" category; the other was the only one awarded in the "Novel Approaches" category.

- Robust recognition (in noise and reverberation): New results for one or two microphones recording speech at a distance (3-6 feet) in room with typical conference room noise and reverberation.

- Spoken language in meetings: Meeting Corpus completed, going to the Linguistic Data Consortium for public (international) use; Corpus and related annotations the

basis of widespread activity in US and Europe, forming the largest carefully transcribed public collection of this kind. A new extension of this effort was started in collaboration with the Swiss network IM2.

- Speaker recognition: in cooperation with multi-site collaborators, incorporated prosody and linguistic features along with acoustics, achieving best results yet on common test set.

- Dialog systems: ICSI developed (in cooperation with multi-site partners) the English language version of the SmartKom Mobile demonstration system (SmartKom 3.2). The system – a personal travel information assistant – was then demonstrated during a 3-day period at the International Conference for Spoken Language Systems (ICSLP-2002) in Denver, Colorado.

BERKELEY CENTER FOR THE INFORMATION SOCIETY (BCIS)

- Inauguration of the Center on September 26, 2002.

- Publication of Castells and Himanen book, *The Information Society and the Welfare State : The Finnish Model* (Oxford University Press, 2002).

- A number of initial seminars and talks, including a meeting with the heads of the principal Finnish research organizations.

NETWORKING GROUP

INTERNET ARCHITECTURE

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 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. TCP can introduce arbitrary delay because of its reliability and in-order delivery requirements; thus, these delay-sensitive and loss-tolerant applications use UDP instead. 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. Consequently, we believe that a new protocol is needed, one that combines unreliable datagram delivery with built-in congestion control. We have begun the design of the Datagram Congestion Control Protocol (DCCP) that implements a congestion-controlled, unreliable flow of datagram suitable for use by applications such as streaming media or on-line games. DCCP is currently being standardized in the IETF.

HIGH SPEED TCP: HighSpeed TCP is a proposed 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 have proposed HighSpeed TCP, a modification to TCP's congestion control mechanisms. One attractive quality of HighSpeed TCP is that it would be easy to deploy, only needing to be implemented at the transport data sender.

ROUTER-BASED CONGESTION CONTROL

XCP: The eXplicit Control Protocol, XCP, generalizes the Explicit Congestion Notification proposal (ECN) and introduces the new concept of decoupling utilization control from fairness control. This allows a more flexible and analytically tractable protocol design and opens new avenues for service differentiation. XCP is stable and efficient regardless of the link capacity, the round trip delay, and the number of sources. Extensive packet-level simulations show that XCP outperforms TCP in both conventional and high bandwidth-delay environments. Further, XCP achieves fair bandwidth allocation, high utilization, small standing queue size, and near-zero packet drops, with both steady and highly varying traffic. Additionally, the new protocol does not maintain any per-flow state in routers and requires few CPU cycles per packet, which makes it implementable in high-speed routers.

QUICKSTART: Quick-Start is a proposed mechanism for faster start-up for TCP and other transport protocols in environments with significant available bandwidth. The current TCP protocol starts-up using an initial window of up to four packets, and then the TCP sender uses slow-start to double the sending rate each round-trip time until congestion is encountered. As a result, it may take a number of round-trip times in slow-start before the TCP connection begins to fully use the available bandwidth. Quick-Start is a proposal that would allow TCP connections to start-up faster, by asking the routers along the path to approve a higher initial sending rate. Quick-Start is designed to allow TCP connections to use high initial windows in circumstances when there is significant unused bandwidth along the path, and all of the routers along the path support the Quick-Start Request.

ECN: ECN (Explicit Congestion Notification), standardized in 2001 in RFC 3168, allows routers to use the Congestion Experienced (CE) codepoint in a packet header as an indication of congestion, instead of relying solely on packet drops. However, the work on a protocol does not always end when the paper is published, or when the protocol has been standardized for use in the Internet. The initial deployment of ECN was complicated by the existence of firewalls in the Internet that inappropriately reset a TCP connection upon receiving certain TCP SYN packets, in particular, packets with flags set in the Reserved field of the TCP header. This led to an examination of the use of TCP resets by firewalls in the Internet, and a consideration of the longer-term consequences of this and similar actions as obstacles to the evolution of the Internet infrastructure.

FLOW-TABLE FAIRNESS MECHANISMS: The original Approximate Fair Dropping (AFD) proposal required substantial state. A new design, using only a flow table rather than a shadow buffer, matches the performance of the AFD proposal but requires far less state.

ROUTING

ROUTE-FLAP DAMPING: Route flap damping is a widely deployed mechanism in core routers to limit the widespread propagation of unstable BGP routing information. Originally designed to suppress route changes caused by link flaps, flap damping attempts to distinguish persistently unstable routes from routes that occasionally fail. It is considered to be a major contributor to the stability of the Internet routing system. Surprisingly, it turns out that route flap damping can significantly exacerbate the convergence times of relatively stable routes. For example, a route to a prefix that is withdrawn exactly once and re-announced can be suppressed for up to an hour (using the current RIPE recommended damping parameters). This abnormal behavior fundamentally arises from the interaction of flap damping with BGP path exploration during route withdrawal. We studied this interaction using a simple analytical model and analyzed the impact of various BGP parameters on its occurrence using simulations. We also developed a preliminary proposal to modify a route flap damping scheme that removes the undesired interaction in all the topologies we studied.

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 a 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.

NOVEL APPROACHES

ROLE-BASED ARCHITECTURES: Questioning whether layering is still an adequate foundation for networking architectures, this project investigates non-layered approaches to the design and implementation of network protocols. The goals are greater flexibility and control with fewer feature interaction problems. We propose a specific non-layered paradigm called role-based architecture.

DESIGNING ROBUST PROTOCOLS: Robustness has long been a central design goal of the Internet. Much of the initial effort towards robustness focused on the "fail-stop" model, where node failures are complete and easily detectable by other nodes. The Internet is quite robust against such failures, routinely surviving various catastrophes with only limited outages. This robustness is largely due to the widespread belief in a set of guidelines for critical design decisions such as where to initiate recovery and how to maintain state. However, the Internet remains extremely vulnerable to more arbitrary failures where, through either error or malice, a node issues syntactically correct responses that are not semantically correct. Such failures, some as simple as misconfigured routing state, can seriously undermine the functioning of the Internet. With the Internet playing such a central role in the global telecommunications infrastructure, this level of vulnerability is no longer acceptable. In this project we argue that to make the Internet more robust to these kinds of arbitrary failures, we need to change the way we design network protocols. To this end, we propose a set of six design guidelines for improving the network protocol design. These guidelines emerged from a study of past examples of failures, and determining what could have been done to prevent the problem from occurring in the first place. The unifying theme behind the various guidelines is that we need to design protocols more defensively, expecting malicious attack, misimplementation, and misconfiguration at every turn.

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 have begun a project to broaden discussion within the research community about the models we use, with the goal of a more agreed-upon set of best modeling practices. Our underlying assumptions are that models should be as simple as possible for the specific purpose but no simpler, and that, as a result, models need to be specific to the research questions under investigation.

NIMI: NIMI (National Internet Measurement Infrastructure) is an on-going project to develop and deploy a system for facilitating coordinated measurement from a number of points around the Internet. The work currently focuses primarily on refining the authorization, security, and resource control mechanisms.

EFFECT OF 9-11: A study by the Computer Science and Telecommunications Board of the National Research Council on the effect on the Internet of the September 11, 2001, terrorist attacks, especially the destruction of the World Trade Center towers.

PKTD: Administrators can be highly reluctant to run foreign measurement tools on their hosts because such tools frequently require privileged execution in order to transmit customized measurement packets and/or to passively capture network traffic. Also, the administrators lack mechanisms to control the rate, duration, type, destination, and contents of traffic generated by the measurements. We are working on developing "pktd", a packet capture and injection multiplexer daemon that provides controlled, fine-grained access to the network device. On systems running pktd, client measurement tools are not given direct access to the network device, but instead are obliged to request access via the daemon, which then provides a single point of privilege and control.

ACTIVE MAPPING: A critical problem faced by network intrusion detection systems is that of *ambiguity* ---the NIDS does not always know what traffic reaches a given host nor how that host will interpret the traffic. This problem is fundamental to passive monitors; without knowledge of the context in which it is running (i.e., the network topology and the TCP/IP policies of the hosts it is protecting). It is in general impossible for a NIDS to accurately resolve TCP/IP-based ambiguities, and these ambiguities can then be exploited by attackers to confuse or evade detection by the NIDS (indeed, toolkits to facilitate such evasion already exist). This project explores a way to resolve such ambiguities by proactively building profiles of the network topology and the TCP/IP policies of hosts on the network by sending specially crafted packets to each host. A major advantage of Active Mapping over previous work is that it does not require any interception or modification of the traffic stream.

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 develops 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.

ADDRESS STRUCTURE: This project investigates the structure of addresses contained in IP traffic. The work has implications for algorithms that deal with IP address aggregates, such as routing lookups and aggregate-based congestion control. One of the main interesting findings to date is that address structures are well modeled using a multifractal construction with only two parameters.

FLOW RATES: This project analyzes the distribution of the rates at which flows transmit data, and the causes of these rates. The heart of the analysis is the development of a tool, T-RAT, to analyze packet-level TCP dynamics. In our study, the most frequent reasons for the rates achieved by connections appear to be network congestion and receiver window limits.

PEER-TO-PEER SYSTEMS

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.

INTERNET INDIRECTION INFRASTRUCTURE: Indirection plays a fundamental role in today's Internet. Existing solutions to provide mobility, anycast, and multicast at the IP layer, and to provide web caching and server load balancing at the application layer, are based on indirection. Unfortunately, the fact that IP does not provide efficient support for indirection makes it difficult and complex to deploy these solutions. In this project, we propose to replace the point-to-point communication abstraction used in today's networks with a rendezvous-based communication abstraction: instead of explicitly sending a packet to a destination, each packet is associated with an identifier, which is then used by the receiver to get the packet. This decoupling between the sender and the receiver allows rendezvous-based networks to provide natural support for mobility, anycast, and multicast.

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.

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.

REPLICATION IN FILE-SHARING SYSTEMS: The Peer-to-Peer (P2P) architectures that are most prevalent in today's Internet are decentralized and unstructured. Search is blind in that it is independent of the query and is thus not more effective than probing randomly chosen peers. One technique to improve the effectiveness of blind search is to proactively replicate data. We evaluate and compare different replication strategies and reveal interesting structure: Two very common but very different replication strategies -- uniform and proportional -- yield the same average performance on successful queries, and are in fact worse than any replication strategy which lies between them. The optimal strategy lies between the two and can be achieved by simple distributed algorithms. These fundamental results offer a new understanding of replication and show that currently deployed replication strategies are far from optimal and that optimal replication is attainable by protocols that resemble existing ones in simplicity and operation.

SECURITY 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.

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. This project adds a signature-matching engine into the Bro intrusion detection

system, with the key being integration between the engine and Bro's powerful contextual analysis capabilities.

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 on-going 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 TRIGGERS: Many automated network attacks have the form in which an initial connection to a host triggers (upon success) a subsequent connection, either inbound from the attacker to test whether a "back door" has been established, or outbound from the victim to signal to the attacker that the exploit succeeded. This project aims to develop heuristics for detecting such attacks.

SECURITY FOR MOBILE AGENTS: Recent work has shown that several cryptographic protocols for the protection of free-roaming mobile agents are vulnerable by means of protocol interleaving attacks. We have developed equivalent protocols meant to be robust against this type of attack. Moreover, it describes the required processes and data structures at a level of detail that can be translated to an implementation in a straightforward way. Our aim is to demonstrate how cryptographic processing can be implemented transparently for agent programmers, thereby reducing the risks of human error in (secure) mobile agent programming.

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 its first alpha release in December.

CLICK MODULAR ROUTER: Work on the Click modular router continues, both within the context of the XORP project and independently. Ongoing efforts focus on improving stability. Also we have greatly improved the usefulness of Click for trace analysis by developing a large set of reusable analysis components.

SENSORNETS

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.

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 and Mark Handley are currently members of the Internet Architecture Board (IAB), which is a technical advisory board to the Internet Society and the IETF.

REFERENCES

PAPERS

[1] A. Akella, R. Karp, C. Papadimitriou, S. Seshan and S. Shenker. Selfish Behavior and Stability of the Internet: A Game-Theoretic Analysis of TCP. In *Proceedings ACM SIGCOMM 2002*.

[2] T. Anderson, S. Shenker, I. Stoica, and D. Wetherall. Design Guidelines for Robust Internet Protocols. In *First Workshop on Hot Topics in Networks (HotNets-I)* October 2002.

[3] Walter Binder and Volker Roth. Secure mobile agent systems using Java - where are we heading? In *Proceedings. 17th ACM Symposium on Applied Computing, Special Track on Agents, Interactions, Mobility, and Systems (SAC/AIMS)*, Madrid, Spain, March 2002. ACM.

[4] B. Braden, T. Faber, and M. Handley. From Protocol Stack to Protocol Heap: Role-based Architecture. In *First Workshop on Hot Topics in Networks (HotNets-I)* October 2002.

[5] Q. Chen, H. Chang, R Govindan, S. Jamin, S. Shenker, W. Willinger. The Origin of Power Laws in Internet Topologies Revisited. In *Proceedings of* . *IEEE Infocom 2002*.

[6] E. Cohen and S. Shenker. Replication Strategies in Unstructured Peer-to-Peer Networks. In *Proceedings ACM SIGCOMM 2002*.

[7] D. Donoho, A. G. Flesia, U. Shankar, V. Paxson, J. Coit, and S. Staniford. Multiscale Stepping-Stone Detection: Detecting Pairs of Jittered Interactive Streams by Exploiting Maximum Tolerable Delay. In *Proceedings of Fifth International Symposium on Recent Advances in Intrusion Detection (RAID) 2002*.

[8] J. Feigenbaum, C. Papadimitriou, R. Sami, and S. Shenker, A BGP-based Mechanism for Lowest-Cost Routing. In *Proceedings of the Twenty-First Annual ACM Symposium on Principles of Distributed Computing (PODC) 2002*.

[9] S. Floyd and E. Kohler. Internet Research Needs Better Models. *Workshop Record of the First Workshop on Hot Topics in Networks (HotNets-I)*, Princeton, New Jersey, October 2002.

[10] R. Govindan and V. Paxson. Estimating Router ICMP Generation Delays. In *Proceedings of Passive & Active Measurement: (PAM) 2002*.

[11] M. Handley, O. Hodson, and E. Kohler. XORP: Open Platforms for Network Research. In *Workshop Record of the First Workshop on Hot Topics in Networks (HotNets-I)* October 2002.

[12] M. Harren, J. M. Hellerstein, R. Huebsch, B. T. Loo, S. Shenker and I. Stoica. Complex Queries in DHT-based Peer-to-Peer Networks. In *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS) 2002*, March 2002.

[13] D. Katabi, M. Handley, and C. Rohrs. Internet Congestion Control for High Bandwidth-Delay Product Networks. In *Proceedings of the ACM SIGCOMM, August 2002*.

[14] E. Kohler, J. Li, V. Paxson and S. Shenker. Observed Structure of Addresses in IP Traffic. In *Proceedings. of. ACM SIGCOMM Internet Measurement Workshop* November 2002.

[15] E. Kohler, J. Li, V. Paxson, and S. Shenker. Observed structure of addresses in IP traffic. In *Proceedings of the 2nd Internet Measurement Workshop (IMW) 2002*, Marseille, France, November 2002, pages 253-266.

[16] E. Kohler, R. Morris, and B. Chen. Programming language optimizations for modular router configurations. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X*), San Jose, California, October 2002, pages 251-263.

[17] E. Kohler, R. Morris, and M. Poletto. Modular components for network address translation. In *Proceedings of the 5th International Conference on Open Architectures and Network Programming (OPENARCH) 2002*, New York, New York, June 2002, pages 39-50.

[18] Q. Lv, S. Ratnasamy and S. Shenker. Can Heterogeneity Make Gnutella Scalable? In *Proceedings. of First International Workshop on Peer-to-Peer Systems (IPTPS) 2002*.

[19] R. Mahajan, S. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, and S. Shenker. Controlling High Bandwidth Aggregates in the Network. *Computer Communication Review*, V.32, N.3, July 2002. (Listed in 2001 as "submitted to CCR".)

[20] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz. Route Flap Dampening Exacerbates Internet Routing Convergence. In *Proceedings of the ACM SIGCOMM, August 2002.*

[21] R. Pan, L. Breslau, B. Prabhakar, and S. Shenker. A Flow Table-Based Design to Approximate Fairness. In *Hot Interconnects: 10th Symposium on High Performance Interconnects (Hot-I) 2002.*

[22] C. Partridge, P. Barford, D. Clark, S. Donelan, V. Paxson J. Rexford, M. Vernon, J. Eisenberg, M. Blumenthal, D. Padgham, K. Batch, D. Drake, and J. Briscoe. The Internet Under Crisis Conditions: Learning from September 11. Computer Science and Telecommunications Board, National Research Council, National Academy Press, Washington, D.C., 2002.

[23] Ulrich Pinsdorf and Volker Roth. Mobile Agent Interoperability Patterns and Practice. In *Proceedings 9th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems*, Lund, Sweden, April 2002.

[24] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker. Topologically-Aware Overlay Construction and Server Selection. In *Proceedings of IEEE INFOCOM 2002*, July 2002.

[25] S. Ratnasamy, B. Karp, Li Yin, Fang Yu, D. Estrin, R. Govindan, and S. Shenker. GHT: A Geographic Hash-table for Data-centric Storage in Sensornets. In *First ACM International Workshop on Wireless Sensor Networks and Applications (WSNA),* September 2002.

[26] S. Ratnasamy, I. Stoica and S. Shenker. Routing Algorithms in DHTs: Some Open Questions. In *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS) 2002*, March 2002.

[27] V. Roth. Java Security Architecture and Extensions. *Dr. Dobbs Journal*, 2002(335), April 2002.

[28] V. Roth. Empowering mobile software agents. In N. Suri, editor, *Proceedings. 6th IEEE Mobile Agents Conference,* volume 2535 of *Lecture Notes in Computer Science*, pages 47-63. Springer Verlag, October 2002.

[29] V. Roth and M. Arnold. Improved key management for digital watermark monitoring. *Proceedings of SPIE*, San Jose, CA, USA, January 2002.

[30] S. Shenker, S. Ratnasamy, B. Karp, R. Govindan, and Deborah Estrin.. Data-centric Storage in Sensornets. In *Workshop Record of the First Workshop on Hot Topics in Networks (HotNets-I)* October 2002.

[31] S. Staniford, V. Paxson and N. Weaver. How to 0wn the Internet in Your Spare Time. In *Proceedings of USENIX Security Symposium 2002.*

[32] I. Stoica, D. Adkins, S. Ratnasamy, S. Shenker, S. Surana, S. Zhuang. Internet Indirection Infrastructure. In *Proceedings. of First International Workshop on Peer-to-Peer Systems (IPTPS) 2002*.

[33] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, and S. Surana.. Internet Indirection Infrastructure In *Proceedings. ACM SIGCOMM, August 2002*.

[34] I. Stoica, D. Adkins, S. Ratnasamy, S. Surana, S. Shenker, and S. Zhuang. Routing Algorithms in DHTs: Some Open Questions. *In Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS) 2002*, March 2002.

[35] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker and W. Willinger. Network Topology Generators: Degree-Based vs Structural. In *Proceedings ACM SIGCOMM 2002*.

[36] W. Willinger, R. Govindan, S. Jamin, V. Paxson and S. Shenker. Scaling phenomena in the Internet: Critically examining criticality. In *Proceedings of the National. Academy of Sciences. USA*. Vol. 99, Suppl. 1, 2573-2580, February 19, 2002.

[37] W. Willinger, V. Paxson, R. H. Riedi and M. S. Taqqu. Long-range Dependence and Data Network Traffic. To appear in *Long-range Dependence: Theory and Applications*, P. Doukhan, G. Oppenheim and M. S. Taqqu, eds., Birkhauser, 2002.

[38] M. Zhang, B. Karp, S. Floyd, and L. Peterson. RR-TCP: A Reordering-Robust TCP with DSACK. ICSI Technical Report TR-02-006, Berkeley, CA, July 2002.

[39] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker. On the Characteristics and Origins of Internet Flow Rates. In *Proceedings ACM SIGCOMM 2002*.

INTERNET DRAFTS

[40] S. Floyd. HighSpeed TCP for Large Congestion Windows. Internet draft draft-floyd-tcp-highspeed-01.txt, work in progress, August 2002.

[41] S. Floyd. Limited Slow-Start for TCP with Large Congestion Windows. Internet draft draft-floyd-tcp-slowstart-01.txt, work in progress, August 2002.

[42] S. Floyd and E. Kohler. Profile for DCCP Congestion Control ID 2: TCP-like Congestion Control, Internet draft draft-ietf-dccp-ccid2-00.txt, work in progress, October 2002.

[43] S. Floyd, M. Handley, and E. Kohler. Problem Statement for DCCP. Internet draft draft-ietf-dccp-problem-00.txt, work in progress, October 2002.

[44] M. Handley, J. Pahdye, S. Floyd, and J. Widmer. TCP Friendly Rate Control (TFRC): Protocol Specification. Internet draft draft-ietf-tsvwg-tfrc-05.txt, work in progress, October 2002. Approved for Proposed Standard. A. Jain and S. Floyd, Quick-Start for TCP and IP. Internet-draft draft-amit-quick-start-02.txt, work in progress, October 2002.

[45] E. Kohler, M. Handley, S. Floyd, and J. Padhye. Datagram Congestion Control Protocol (DCCP). Internet draft draft-ietf-dccp-spec-00.txt, work in progress, October 2002.

[46] J. Padhye, S. Floyd, and E. Kohler. Profile for DCCP Congestion Control ID 3: TFRC Congestion Control. Internet draft draft-ietf-dccp-ccid3-00.txt, work in progress, October 2002.

[47] K. Wehrle, Kathleen Nichols, and Roland Bless. A Lower Effort Per-Domain Behavior for Differentiated Services. Internet-Draft draft-bless-diffserv-pdb-le-01, November 2002

[48] K. Wehrle and R.. Bless: IP Multicast in Differentiated Services Networks. Internet-Draft draft-bless-diffserv-multicast-05.txt, November 2002

RFCs

[49] M. Allman, S. Floyd, and C. Partridge. Increasing TCP's Initial Window, RFC 3390, PROPOSED STANDARD, October 2002. Obsoletes RFC 2414. S. Floyd (Editor), General Architectural and Policy Considerations, RFC 3426, INFORMATIONAL, November 2002.

[50] S. Floyd. Inappropriate TCP Resets Considered Harmful, RFC 3360, BEST CURRENT PRACTICE, August 2002.

[51] S. Floyd and L. Daigle (Eds). IAB Architectural and Policy Considerations for Open Pluggable Edge Services, RFC 3238, INFORMATIONAL, January 2002.

IN SUBMISSION

[52] N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, "Large Scale Malicious Code: A Research Agenda."

[53] R Sommer and V. Paxson, "Detecting Network Intruders Using Contextual Signatures."

[54] U. Shankar and V. Paxson, "Active Mapping: Resisting NIDS Evasion Without Altering Traffic."

[55] J. M. Gonzalez and V. Paxson, "pktd: A Packet Capture and Injection Daemon."

ALGORITHMS

During 2002 the Algorithms Group conducted research on combinatorial algorithms, information networks and computational genomics. Within the area of combinatorial algorithms we made path-breaking progress on approximation algorithms for NP-hard problems; other prominent themes were probabilistic combinatorics, combinatorial games and geometric methods. The work on information networks was conducted jointly with the ICIR Group and focused on Internet congestion and resource allocation in peer-to-peer networks. Particularly striking was an upsurge in the computational biology activity, with a wide range of genomics-related activities and a special emphasis on two key themes in functional genomics: haplotyping and analysis of transcriptional regulation. The members of the group were Group Leader Richard Karp (combinatorial algorithms, information networks, computational biology), Board Chair Elwyn Berlekamp (combinatorial game theory), postdoctoral fellows Ernst Althaus (combinatorial algorithms), Eran Halperin (approximation algorithms, computational biology), Robert Krauthgamer (approximation algorithms, metric space embeddings), Roded Sharan (computational biology), Matthias Westermann (distributed computing), graduate student Eric Xing (computational biology, machine learning), and undergraduate student Kushal Chakrabarti (computational biology). The Group ran a weekly seminar throughout the year, with many participants from ICSI and the Berkeley campus.

COMBINATORIAL ALGORITHMS

APPROXIMATION ALGORITHMS

In this field we consider polynomial-time algorithms to obtain approximate algorithms for optimization problems that are NP-hard, and therefore presumed to be intractable. A key concept is the approximation ratio f(n) of a polynomial-time algorithm, defined as the maximum, over problem instances of size n, of the ratio between the cost of the solution the algorithm produces and the cost of an optimal solution.

POLYLOGARITHMIC INAPPROXIMABILITY

Several natural combinatorial optimization problems (such as Bandwidth, Pathwidth, Group-Steiner-Tree, Job-Shop Scheduling, and Min-Bisection) are known to have algorithms achieving an approximation ratio that is polylogarithmic in the input size. However, none of these problems were known to have a hardness of approximation result that excludes the possibility of a logarithmic approximation.

The paper [21] gives an $\Omega(\log^2 k)$ lower bound on the integrality ratio of a linear programming relaxation of the Group-Steiner-Tree problem, where k denotes the number of groups. This yields the first polylogarithmic integrality ratio for a natural relaxation of a problem.

In [20] we build on this integrality ratio to show a polylogarithmic hardness of approximation result for the Group-Steiner-Tree problem. We prove that, for every fixed $\varepsilon > 0$, the problem admits no efficient $\log^{2-\varepsilon}(k)$-approximation, where k denotes the number of groups (or, alternatively, the input size), unless NP has quasi-polynomial Las

Vegas algorithms. Both results hold even for input graphs that are trees, in which case they are nearly tight with the algorithmic upper bounds of [14].

CONSTANT FACTOR APPROXIMATION OF VERTEX CUTS IN PLANAR GRAPHS

Graph partitioning problems have many application both in theory and in practice. In [5] we devise the first constant factor approximation algorithm for minimum quotient vertex cuts in planar graphs. This algorithm yields the first constant factor pseudo-approximation for vertex separators in planar graphs.

Our technical contribution is two-fold. First, we prove a structural theorem for vertex cuts in planar graphs, showing the existence of a near-optimal vertex cut whose high-level structure is that of a bounded-depth tree. Second, we develop an algorithm for optimizing over such complex structures, whose running time depends (exponentially) not on the size of the structure, but only on its depth.

HARDNESS OF APPROXIMATION FOR VERTEX CONNECTIVITY NETWORK DESIGN PROBLEMS

In the survivable network design problem (SNDP) the goal is to find a minimum-cost spanning subgraph satisfying certain connectivity requirements. We study the vertex-connectivity variant of SNDP in which the input specifies, for each pair of vertices, a required number of vertex-disjoint paths connecting them.

In [28] we give the first strong lower bound on the approximability of SNDP, showing that the problem admits no efficient $2^{\log^{1-\varepsilon}(n)}$-ratio approximation for any fixed $\varepsilon > 0$ unless $NP \subseteq DTIME(n^{polylog(n)})$. In contrast, the edge-connectivity variant of the problem has a 2-approximation algorithm. We also show hardness of approximation results for several important special cases of SDDP and the first lower bound on the approximability of the related classical NP-hard problem of augmenting the connectivity of a graph using edges from a given set.

IMPROVED APPROXIMATION ALGORITHMS FOR THE PARTIAL VERTEX COVER PROBLEM

The partial vertex cover problem is a generalization of the vertex cover problem: given an undirected graph G = (V, E) and an integer k, we wish to choose a minimum number of vertices such that at least k edges are covered. Just as for vertex cover, 2-approximation algorithms are known for this problem. The previous best approximation ratio for partial vertex cover, when parametrized by the maximum degree d of G, is $2 - \Theta(1/d)$. In [22] we improve on this by presenting a $2 - \Theta\left(\frac{\ln \ln d}{\ln d}\right)$-approximation algorithm using semi-definite programming, matching the current best bound for vertex cover. Our algorithm uses a new rounding technique, which involves a delicate probabilistic analysis.

OTHER ALGORITHMIC RESULTS

Here we describe a number of algorithmic results based on methods from probability, combinatorics, geometry and machine learning.

DISCRETE STOCHASTIC PROCESSES

In [2] we consider a stochastic process executed on a k-dimensional hypercube whose nodes are initially uncovered. At each step, pick a node at random and, if it is uncovered, cover it. Otherwise, if it has an uncovered neighbor, cover a random such neighbor. Else, do nothing. We show that $O(2^k)$ steps suffice to cover all nodes of the hypercube with high probability. The process arises naturally in connection with load balancing in peer-to-peer networks.

In [25] we analyze a certain gambling game which arises in the analysis of adaptive randomized rounding algorithms for combinatorial optimization problems related to multicommodity flow. Our approach gives stronger results than can be obtained using the traditional Hoeffding bounds and martingale tail inequalities.

In [1] we consider a discrete-time stochastic process where, in each round, n balls are tossed into bins, and the balls landing in each bin are coalesced into a single ball. We study the distribution of the number of rounds until only a single ball remains. Our result applies to a generalization of the Wright-Fisher model in population genetics.

THE INTRINSIC DIMENSIONALITY OF GRAPHS

Levin together with Linial, London, and Rabinovich raised the following conjecture. Let $dim(G)$ be the smallest d such that G occurs as a (not necessarily induced) subgraph of the infinite d-dimensional unit grid. The *growth rate* of G, denoted ρ_G, is the minimum ρ such that every ball of radius $r > 1$ in G contains at most r^{ρ} vertices. By simple volume arguments, $Dim(G) = \Omega(\rho_G)$. Levin conjectured that this lower bound is tight, i.e., that $dim(G) = O(\rho_G)$ for every graph G.

In [29] we resolve Levin's conjecture. We show that $dim(G) = O(\rho_G \log \rho_G)$ and this upper bound is tight, disproving the original conjecture. We also show that Levin's conjecture holds for certain families of graphs.

GENERALIZED VARIATIONAL INFERENCE

In [39] we present a generalized variational inference (GVI) algorithm for efficient approximate inference in complex probabilistic models. GVI approximates a complex joint posterior with a simpler distribution factored over a disjoint variable clustering. With no need for model-driven formulation of structural variational inference equations, GVI uses a set of canonical fixed-point equations to iteratively update factors of the approximate distribution, which leads to guaranteed convergence to locally optimum cluster marginals and provides a bound on the approximation error. Under mild technical assumptions satisfied in most models in practical use, the cluster marginals take a simple form that preserves the original independencies among variables in the cluster, hence localizing and simplifying approximate inference. We apply GVI to relational probability models, whose semantic structure is exploited to motivate heuristic variable clusterings. Empirical results suggest that GVI outperforms MCMC with regard to accuracy, efficiency and scalability.

OTHER CONTRIBUTIONS

[6] provides lower bounds on the average-case complexity of resolution methods for proving the unsatisfiability of propositional disjunctive normal form formulas. [26] gives a simple and efficient algorithm for finding frequent tokens in a stream of tokens, an abstraction of the

problems of finding highly frequent database queries and high intensity Internet flows. [19] establishes favorable bounds on the performance of a greedy algorithm for the following set partioning problem: given a finite universe U and a family of sets F, partition U into sets S_i such that each is a subset of some set in the family F, so as to maximize $\sum |S_i| log |S_i|$ a measure of concentration of the partition.

INFORMATION NETWORKS

LOAD BALANCING IN PEER-TO-PEER NETWORKS

[31] addresses the problem of load balancing in P2P systems that provide a distributed hash table (DHT) abstraction. It assumes that the address space is partitioned into regions associated with virtual servers. Each virtual server is assigned to a physical server, and load balancing can be achieved by changing this assignment. We present and compare three simple load-balancing schemes that differ primarily in the amount of information used to decide how to rearrange load. Metrics of performance are the amount of data movement and the degree of agreement between the capacities of physical servers and their loads.

In a complementary approach, [2] studies the problem of dynamically partitioning the address space of a DHT so that the largest and smallest regions assigned to individual servers differ in volume by at most a constant factor. By analyzing a related stochastic process we show that a natural scheme previously studied by simulation satisfies this criterion with high probability.

GAME-THEORETIC ANAYSIS OF TCP

For years, the conventional wisdom has been that the continued stability of the Internet depends on the widespread deployment of ``socially responsible" congestion control. In [3] we seek to answer the following fundamental question: if network end-points behaved in a selfish manner, would the stability of the Internet be endangered? We evaluate the impact of greedy end-point behavior through a game-theoretic analysis of TCP. In this "TCP game" each flow attempts to maximize its throughput by modifying its congestion control behavior. Through simulation and analysis we study whether the network operates efficiently at the Nash equilibria of this game. We find that more traditional versions of TCP can remain stable in the face of greedy end-user behavior, but more recent variations tend to lead to inefficient Nash equilibria.

COMPUTATIONAL GENOMICS

The Algorithms Group studied a wide variety of problems related to understanding the analysis of DNA and protein sequences, the structure of genomes, the expression of genes and the effects of genetic variation on disease, with particular emphasis on haplotyping, analysis of cis-regulation and clustering.

HAPLOTYPING

Each person's genome contains two copies of each chromosome, one inherited from the father and the other from the mother. A person's *genotype* specifies the pair of bases at each site, but does not specify which base occurs on which copy of the chromosome. The sequence of each chromosome separately is called a *haplotype*. The determination of the haplotypes within a population is essential for understanding genetic variation and the

inheritance of complex diseases. The haplotype mapping project, a successor to the human genome project, seeks to determine the common haplotypes in the human population.

Since experimental determination of a person's genotype is less expensive than determining its component haplotypes, algorithms are required for computing haplotypes from genotypes. Two observations aid in this process: first, the human genome contains short blocks within which only a few different haplotypes occur, with the variation mainly due to single-nucleotide polymorphisms (SNPs); second, as suggested by Gusfield, it is reasonable to assume that the common haplotypes within a block have evolved according to a *perfect phylogeny*, in which at most one mutation event has occurred at any site.

In joint work with Eleazar Eskin of Columbia University ([10], [11], [18]) we present a simple and efficient polynomial-time algorithm for inferring haplotypes from the genotypes of a set of individuals assuming a perfect phylogeny. Using a reduction to 2-SAT we extend this algorithm to handle constraints that apply when we have genotypes from both parents and child. We extend these methods to partition the SNPs into blocks and determine the haplotypes within each block, even when there are deviations from perfect phylogeny. We evaluate the method on biological data. Our method predicts the common haplotypes perfectly and has a very low error rate (0.47%) when taking into account the predictions for the uncommon haplotypes. Furthermore, our algorithms can cope with missing data.

In [15] we study the problem of analyzing the haplotype block structure of a population that consists of several unknown subpopulations with distinct block structures. The general problem that we study is NP_hard. We heuristically tackle it using a combination of a dynamic programming algorithm for a restricted case, and simulated annealing. Our algorithm is based on a new scoring function for a block structure, which handles both missing data and data errors. We are in the process of applying this algorithm to simulated and real haplotype data.

In [8] we study a design and optimization problem that occurs when SNPs are to be genotyped using a universal DNA tag array. 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, however, also occur between primers and tags. We examined the problem of identifying the most economical experimental design that avoids cross-hybridization. This translates into the problem of covering the vertices of a bipartite graph by a minimum number of balanced subgraphs of maximum degree 1. We present and evaluate approximation algorithms and heuristic algorithms for this problem.

TRANSCRIPTIONAL REGULATION

The biologist Eric Davidson has described the regulation of transcription in the following terms: Regulatory interactions mandated by circuitry encoded in the genome determine whether each gene is expressed in each cell, throughout developmental space and time, and, if so, at what amplitude. The working parts of the genome are the genes and their *cis*-regulatory elements. The internal architecture of a *cis*-regulatory element enables it to process the various inputs it receives, resolving these inputs into a single output. A *transcription factor* is a protein which displays a high specificity for a particular *cis*-regulatory DNA sequence, and which performs some function that affects transcriptional output. These sequences are often grouped into regulatory *modules* strung out in the DNA flanking the gene or in its introns.

We have developed algorithms for identifying the sequence motifs that characterize transcription factor binding sites, and for finding *cis*-regulatory modules.

MOTIF FINDING

The complexity of the global organization and internal structure of motifs in higher eukaryotic organisms raises significant challenges for motif detection techniques. To achieve successful *de novo* motif detection it is necessary to model the complex dependencies within and among motifs and incorporate biological prior knowledge. In [37] and [38] we present LOGOS, a principled framework for developing, modularizing, extending and computing expressive motif models for complex biopolymer sequence analysis. LOGOS consists of two interacting submodels: HMDM, a local alignment model capturing prior biological knowledge and positional dependence within motifs; and HMM, a global motif distribution model of frequencies and dependencies of motif occurrences. Model parameters can be fit using training motifs within an empirical Bayesian framework. A variational EM algorithm is developed for *de novo* motif detection. LOGOS improves over existing models that ignore biological priors and dependencies in motif structures and motif occurrences, and demonstrates superior performance on both semi-realistic test data and *cis* regulatory sequences from yeast and Drosophila.

REGULATORY NETWORKS

Dissection of regulatory networks that control gene transcription is one of the greatest challenges in functional genomics. By utilizing human genomic sequences, gene expression data and models for binding sites of known transcription factors, we demonstrated in [13] 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 demonstrated only in prokaryotes and low eukaryotes. We developed computational methods for identifying putative binding sites of transcription factors and 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.

In [34] we study the interplay among several transcription factors in regulating genes. A recurrent combination of binding sites for transcription factors that cooperate in the regulation of genes is called a *cis*-regulatory module (CRM). We established a framework for finding CRMs and scoring their statistical significance. Our method builds on new statistical measures for the enrichment of binding sites and combinations of binding sites in a given set of genes. We have identified such combinations that are enriched in cell cycle genes and in stress-related genes, and have shown that these combinations in some cases exhibit significant similarities in their gene expression patterns or functional annotations.

CLUSTERING

GENE EXPRESSION ANALYSIS

In [33] we present a new Java-based graphical tool called EXPANDER (EXPression ANalyzer and DisplayER), for gene expression analysis and visualization. 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-cluster evaluations via fitness scores and functional enrichment tests.

EVALUATION OF CLUSTERING

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. In [15] we developed a novel statistically-based method for assessing a clustering solution according to prior biological knowledge. Our method can be applied to compare 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 a between-groups variance estimator to a within-groups 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 used our method to evaluate popular clustering methods on a yeast cell-cycle gene expression dataset.

DISTANCE METRIC LEARNING

Many algorithms rely critically on being given a good metric over their inputs. For instance, data can often be clustered in many plausible ways, corresponding to different measures of similarity between the objects being clustered. It is desirable to provide a more systematic way for users to indicate what they consider similar. For instance, we may ask them to provide examples. In [36] we present an algorithm that, given examples of similar (and, if desired, dissimilar) pairs of points in R^n, learns a metric that respects these relationships. Our method is based on posing metric learning as a convex optimization problem, which allows us to give efficient algorithms that avoid local optima. We also demonstrate empirically that the learned metrics can be used to significantly improve clustering performance.

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. In [32] we considered the problem of making the fewest changes to the edge set of an input graph so that it becomes a cluster graph. We showed that several versions of this problem are NP-hard and even hard to approximate, and also exhibited special cases solvable in polynomial time.

DISCOVERING LOCAL STRUCTURE IN GENE EXPRESSION DATA

The paper [7] concerns the discovery of patterns in gene expression matrix, in which each element gives the expression level of a given gene in a given experiment. Most existing methods for pattern discovery in such matrices are based on clustering all the rows, or all the columns of such a matrix. Instead we focus simultaneously on a subset G of the rows and a

subset *T* of the columns. Specifically, we look for *order-preserving submatrices*, in which the expression levels of all genes induce the same linear order of the experiments. Such a pattern might arise, for example, if the experiments in *T* represent different stages of a cellular process, and the expression levels of all genes in *G* vary according to stages in the same way. Application of the method to breast cancer data seems to reveal significant local patterns.

DETECTING PROTEIN SEQUENCE CONSERVATION VIA METRIC EMBEDDING

Comparing two protein databases is a fundamental task in bio-sequence annotation. Given two databases, one must find all pairs of proteins that align with high score under a biologically meaningful substitution score matrix. Distance-based approaches to this problem map each peptide in the database to a point in a metric space, such that peptides aligning with higher scores are mapped to closer points. This work proposes a new distance mapping for peptides that permits efficient similarity search. We first propose a new distance function on peptides derived from a given score matrix. We then use semi-definite programming to map peptides to bit vectors such that the distance between the peptides is closely approximated by the Hamming distance between their corresponding bit vectors. We combine these two results with an existing algorithm to produce an improved distance-based algorithm for proteomic comparison, which exhibits sensitivity within 5% of that of its predecessor and runs eight times faster.

OTHER TOPICS

[24] presents a survey of computational genomics for mathematicians. [4] applies integer programming techniques to the multiple alignment problem. [9] describes a novel approach to the end game of sequence assembly based on constructing a restriction scaffold using complete digestion of the target sequence with several restriction enzymes.

RESEARCH IN COMBINATORIAL GAMES

Bill Fraser completed his doctoral thesis [12] on programs which analyze and solve Go endgame problems. He developed and implemented several novel algorithms for dealing with loopy positions involving kos and superkos. He also developed Go endgame software which uses combinatorial game theory techniques. The main idea is a "divide and conquer" strategy which makes it possible to analyze each of several regions of the board independently, using thermography, and then combine these results into an overall strategy. Although Fraser's software was successful at solving problems in books, attempts to use it to help humans analyze real professional endgames encountered new obstacles. This led to work on a new program which will feature "Autoregion" a set of tools to help the user partition the board into its separate battles, and to track how these regions evolve as the game is played forward or backed up to earlier positions. In early 2003, Brian Carnes launched this new effort with the assistance of Bill Spight, an expert in both mathematical Go and popular Go. There was also great progress in understanding and analyzing a classical children's game called "Fox and Geese". This game is traditionally played on an 8x8 checkerboard. One player has four small checkers, called "geese"; the other has a single king-sized checker, called the "Fox". The geese are allowed only to move forward; the fox can move either forward or back. No jumping is allowed. The player who gets the last move is the winner. When successful, the fox usually wins by escaping to a position behind the geese, where it can stall out the game until all geese reach their finish line and have no more

possible moves. When successful, the geese win only by trapping the fox. It has long been known that Geese can win from the conventional starting position on the 8x8 board.

But many other questions about this game were unanswered until quite recently:

Q: What happens on a larger n x 8 board?
A: Geese can still win.

Q: How big is the geese's advantage on large boards?
A: $1 + 2^{-k}$ moves, where k grows with n.

Q: What does that mean?
A: Games like this can be added and subtracted. To play the difference of two games, on two boards concurrently, each player controls four geese of his color on one board and one fox of his color on the other board. At each turn, he can move any one of his pieces that he chooses. The game ends when one player cannot move on either board. Sometimes the game may draw by going into an infinite loop, such as when both foxes escape. But if the game-theoretic values add up to $1 + 2^{-k} - 1 - 2^{-k} = 0$, then whichever player goes second can win. If the result is positive, Black can win, or if negative, White can win.

Q: If both flocks of geese start on their back rows, does the outcome with perfect play depend on the initial position of the Fox?
A: Yes, on infinitely many (but not all) sizes of boards.

Q: How big is the geese's advantage on small boards?
A: On the 7x8 board, against most positions of the Fox, it is 2 + OVER.

Q: What does that mean?
A: OVER is a very precise, well-defined, well-understood positive infinitesimal with fascinating properties. What looks like its negative image is called UNDER. However, OVER + UNDER = AROUND. Unlike 0, which indicates a win for the second player, AROUND is a game whose outcome is a draw. But it is still infinitesimal. If we add an arbitrarily large number of copies of AROUND to another sum of games whose total value is a positive number, and play the entire sum of all these game, Black could win.

Q: Does this work shed any new light on any other games?
A: YES. Many of the same abstract values appear in Go, Fox-and-Geese, Dots-and-Boxes, and many other games. Furthermore, in early 2003 this work provided the first serious testbed for a powerful new software toolkit being developed by Aaron Seigel. Seigel has implemented a significantly larger subset of mathematical game theory than ever before, and his work is expected to have wide impact next summer when David Wolfe, the author of the toolkit that has been the mainstay of the game research community for the past decade, will help to move all of his users over to the Seigel.

REFERENCES

[1] I. Adler, H.S. Ahn, R.M. Karp, and S.M. Ross. Coalescing times for IID random variables. To appear in *Random Structures and Algorithms* (2003).

[2] M. Adler, E. Halperin, R. Karp, and V. Vazirani. A stochastic process on the hypercube with applications to peer-to-peer networks. Submitted to *Thirty-fifth Annual ACM Symposium on Theory of Computing (STOC2003).*

[3] A. Akella, R.M. Karp, S. Seshan, S. Shenker, and C. Papadimitriou. Selfish behavior and stability of the Internet: a game-theoretic analysis of TCP. In *Proceedings ACM SIGCOMM 2002*.

[4] E. Althaus, A. Caprara, H-P Lenhof, K. Reinert. Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics. In *ECCB*, 4-16 2002.

[5] E. Amir, R. Krauthgamer, and S. Rao. Constant-factor approximation of vertex-cuts in planar graphs. Manuscript 2002.

[6] P. Beame, R.M. Karp, T. Pitassi, and M. Saks. The efficiency of resolution and Davis_Putnamn procedures. In *SIAM J. Comp*. 31(4) 1048-1075 2002.

[7] A. Ben-Dor, B. Chor, R.M. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order preserving submatrix problem. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* 2002.

[8] A. Ben-Dor, T. Hartman, B. Schwikowski, R. Sharan, and Z. Yakhini. Towards optimally multiplexed applications of universal DNA tag systems. To appear in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB)* 2003.

[9] A. Ben-Dor, R.M. Karp, B. Schwikowski, and R. Shamir. The restriction scaffold problem. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* 2002.

[10] E. Eskin, E. Halperin, and R.M. Karp. Large-scale reconstruction of haplotype structure via perfect phylogeny. Technical Report UCB/CSD 2-1196 August, 2002. Also submitted to *Journal of Bioinformatics and Computational Biology*.

[11] E. Eskin, E. Halperin, and R.M. Karp. Large-scale reconstruction of haplotypes from genotype data. To appear in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB)* 2003.

[12] W. Fraser. Computer-assisted thermography of Go endgames. Doctoral dissertation 2002.

[13] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide in-silico identification of transcriptional regulators controlling cell cycle in human cells. To appear in *Genome Research* 2003.

[14] N. Garg, G. Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the group Steiner tree problem. *J. Algorithms* 37(1):66-84, 2000.

[15] I. Gat-Viks, R. Sharan, and R. Shamir. How good is a clustering solution? A. statistical method to score solutions by their biological relevance. (In preparation).

[16] R. Ghandi, E. Halperin, S. Khuller, G. Kortsarz, and A. Srinivasan. Improved bounds for the vertex cover with hard capacities. Work in progress 2002.

[17] E. Halperin, J. Buhler, R.M. Karp, R. Krauthgamer, and B. Westover. Detecting protein conservation via metric embeddings. Submitted to *Intelligent Systems in Molecular Biology* 2003.

[18] E. Halperin and E. Eskin. Large-scale recovery of haplotypes from genotype data using imperfect phylogeny. Technical report no. UCB/CSD-1-1196, August, 2002. Submitted to *Bioinformatics*.

[19] E. Halperin and R.M. Karp. The entropy of the greedy set cover algorithm. In preparation.

[20] E. Halperin and R. Krauthgamer. Polylogarithmic inapproximability. Manuscript 2002.

[21] E. Halperin, G. Kortsarz, R. Krauthgamer, A. Srinivasan, and N. Wang. Integrality ratio for group Steiner trees and directed Steiner trees. To appear in *Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms* 2003.

[22] E. Halperin and R. Srinivasan. Improved approximation algorithms for the partial vertex cover problem. In *5th International Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX*), 185-199, Springer 2002.

[23] M. Handley, R.M. Karp, S. Ratnasamy, and S. Shenker. Topologically-aware overlay construction and server selection. In *Proceedings. INFOCOM* 2000.

[24] R.M. Karp. Mathematical challenges from genomics and molecular biology. In *Notices of the American Mathematical Society*. 49(5) 544-553 .2002.

[25] R.M. Karp and C. Kenyon. A gambling game and its application to the analysis of adaptive randomized rounding. Manuscript 2003.

[26] R.M. Karp, S. Shenker, and C. Papadimitriou. A simple algorithm for finding frequent elements in streams and bags. To appear in *ACM Transactions on Database Systems* 2003.

[27] G. Kimmel, R. Sharan, and R. Shamir. Discovering population blocks from incomplete SNP data with errors. Submitted to *Intelligent Systems in Molecular Biology* 2003.

[28] G. Kortsarz, R. Krauthgamer, and J.R. Lee. Hardness of approximation for vertex-connectivity network design problems. In *5th International Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX)*, 185-199, Springer 2002.

[29] R. Krauthgamer and J.R. Lee. The intrinsic dimensionality of graphs. Manuscript 2002.

[30] H. Racke, C. Sohler, M. Westermann. Online scheduling for sorting buffers. *European Symposium on Algorithms*, 820-832 2002.

[31] A. Rao, K. Lakshminararayanan, S. Surana, R.M. Karp, and I. Stoica. Load balancing in structured P2P systems. In *2nd International Workshop on Peer-to-Peer Systems* 2003.

[32] R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. Submitted to *Discrete Applied Mathematics* 2002.

[33] R. Sharan, A. Maron-Katz, and R. Shamir. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Submitted to *Bioinformatics* 2002.

[34] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R.M. Karp. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Submitted to *Intelligent Systems in Molecular Biology* 2003.

[35] M. Westermann. Distributed caching independent of the network size. *Symposium on Parallel Algorithms and Architecture*s. 31-40 2002.

[36] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. To appear in *Neural Information Processing Systems* 2002.

[37] Eric P. Xing, Michael I. Jordan, Richard M. Karp, and Stuart Russell. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. To appear in *Neural Information Processing Systems* 2002.

[38] Eric P. Xing, Richard M. Karp, and Michael I. Jordan. A modularprobabilistic model for de novo motif detection. Submitted to *Intelligent Systems for Molecular Biology* 2003.

[39] Eric P. Xing and Stuart Russell. On generalized variational inference, with application to relational probability models. Submitted to *International Joint Conference on Artificial Intelligence* 2003.

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.

LANGUAGE LEARNING AND USE

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. Many people will not be in control of the devices in their own environments. In addition, ever widening aspects of society, from education to employment, depend upon everyone interacting with computational systems. 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 machines, and more especially the interactions among the machines, are getting to be so complicated and autonomous, and yet also so intimately involved in the lives of the human users, that they (the machines) have to be able to take part in a kind of social life. 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 (Bailey et al. 97) and of embodied metaphorical reasoning (Narayan 97). About two years ago, we extended our efforts on modeling child language acquisition from individual words and phrases (Regier 96, Bailey 97) 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 (Feldman

98) 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 (Murata, 1989). As discussed below, x-schemas cleanly capture sequentiality, concurrency and event-based asynchronous control and with our extensions they also model the hierarchy and parameterization needed for action semantics.

Our goal is to demonstrate that unifying two powerful linguistic theories, embodied semantics and construction grammar, together with powerful computational techniques, can provide a qualitative 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 to producing a pilot version of the integrated language understanding system [5].

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 an international workshop on Scalable Natural Language Understanding Systems, held in Heidelberg.

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 getting underway; 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 grand competition in the Aquaint program of the U.S. Defense

Department ARDA organization. The group is teaming 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.

The core NTL computational question is finding 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. He and Nancy Chang developed a formal notation for an embodied version of Construction Grammar, which plays a crucial role in a larger, simulation-based language understanding system. They 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. Nancy Chang 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 (Embodied Construction Grammar, devised in collaboration with Ben Bergen). 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 2002 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, Eva Mok, Shweta Narayan, Keith Sanders, 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.

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 [8], 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 [9, 1]. 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.

FRAMENET PROJECT

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 phrase, 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 element**, 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 be automatically analyzed 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 (3) 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.

Our three Co-PIs are developing different NLP applications that take advantage of the strengths of the FrameNet lexicon, and the lexicon is concurrently growing in coverage and adding richer semantic representations, making it better suited for such applications.

+ Co-PI Dan Jurafsky of U Colorado is directing a group using FrameNet for semantic parsing of texts for open-domain Question-Answering, based on mappings of frame-specific roles to high-level thematic roles. Staff at ICSI have collaborated in developing these mappings.

+ Co-PI Srini Narayanan (now at ICSI) is directing a study of information extraction from newspaper accounts of crimes based on texts about criminal acts, arrest, arraignment, court trial, adjudication, and punishment. FrameNet staff have worked on developing a set of relations between the high-level criminal process frame and the frames representing each of these subevents; relating, for example, the arrestee in one step to the defendant in a later step and (perhaps) to the prisoner in a still later step.

+ Co-PI Mark Gawron of San Diego State University is leading research comparing the language of criminal justice in Japanese to the analogous vocabulary in English, for the testing of a machine translation program based on frame descriptions for the crime domain.

A Japanese system for browsing the FrameNet data, maintained at Senshu University by Hiroaki Sato, regularly updated to include new data as the work progresses, has been demonstrated at lexicography research centers in Asia and England.

FrameNet staff have given presentations at various high-profile conferences in the past year, in San Francisco (Linguistic Society of America), Edmonton (American Association of Artificial Intelligence), Las Palmas (Language Resources and Evaluation Conference), and Taipei (COLING), and have offered demos of FrameNet data and tools at Stanford University, the University of Toronto, and Boeing Aircraft. Details can be found at http://www.icsi.berkeley.edu/~framenet/.

CONTACTS AND COLLABORATION

In addition to use by our collaborators in Colorado and San Diego, inquiries about the FrameNet data have been received from a variety of academic and commercial sites. One academic researcher, Joseph Rosenzweig, used our word lists to produce a list of more than 2,000 related terms which he sent us as suggested additions to our frames.

Co-PI Srini Narayanan, then working at SRI, has also extended the FN1 data by adding to the XML attributes in RDF format (using the DAML+OIL conventions) giving information about various relations within the database. Such a scheme should ultimately make the FrameNet data accessible to "smart web" applications.

Our collaboration with the Embodied Construction Grammar group at ICSI continued throughout the year, with members of each group attending meetings of the other. We are trying to ensure that our representations of grammatical constructions and lexical units remain compatible.

CONNECTIONIST MODELING

Over the past year, work on computational modeling has focused on high-level reasoning underlying language understanding and on the formation of episodic memory whereby transient patterns of neural activity representing events and situations are rapidly transformed into persistent neural circuits (memory traces) capable of supporting recognition and recall. The results of these efforts are summarized below.

SHRUTI PROJECT

Lokendra Shastri's work on computational modeling has spanned three different representational and processing tiers of language processing. One tier focuses on high-level reasoning underlying language understanding. The second tier focuses on the formation of episodic memory whereby transient patterns of neural activity representing events and situations are rapidly transformed into persistent neural circuits (memory traces) capable of supporting recognition and recall. The third modeling effort concerns the extraction of syllabic segments from spontaneous and noisy speech. The results of the three efforts are summarized below.

A NEURALLY MOTIVATED MODEL OF REASONING, DECISION MAKING, AND ACTING

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 demonstrates 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.

Over the past year, further progress was made in developing connectionist control structures for selecting among competing actions/plans and implementing a Java-based simulator for SHRUTI. In particular, the SHRUTI progress was made in applying SHRUTI 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 as the target domain for simulation. A simplified SHRUTI knowledge base (KB) to support reasoning and decision-making in this scenario was developed and a more detailed knowledge base is under development.

A number of modifications to the representational machinery and implementation of SHRUTI were carried out. These include:

1) Development of mechanisms for analyzing uncertainty and sensitivity. A three-way breakdown of uncertainty into conflict, incompleteness, and resolution was implemented along with a mechanism for sensitivity analysis.

2) The system's capacity for evidential reasoning was further developed, including the implementation of type-inverse priors.

3) A representation of numerical quantities, and the ability to make comparisons was added and fully integrated with the representation of rules in the Java implementation of SHRUTI.

4) Several improvements to the main graphical user interface (GUI) associated with the SHRUTI simulator were made. These changes facilitate the handling of large domains, allow selective display of node types, and support context-sensitive pop-up menus.

5) The GUI also enables the use of the SHRUTI simulator as an aid in the training of metacognitive/reflective reasoning by allowing a user to identify conflicts and gaps, evaluate assumptions, and select possible actions.

6) Improvements were made in the implementation of multiple (active) instantiations.

Some of the results of the above work are described in [16, 13]. Carter Wendelken is a graduate student and Maximilian Garagnani was a post-doctoral fellow working with Shastri.

A NEURALLY MOTIVATED ARCHITECTURE OF PLANNING

It was shown that a simple connectionist schema acting in concert with two general purpose cognitive faculties, namely, episodic memory and basic perception, can solve a restricted class of planning problems by backchaining from the goal to the current state. In spite of its simple structure, the schema can search for, and execute, plans involving multiple steps. Results of the above work are described in [6].

NEURAL MODELING

There are two related ongoing projects involving detailed neural modeling of important Structure and function, one focusing on memory and the hippocampus and the other on control and the basal ganglia.

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.

SMRITI explicates the representational requirements of encoding events and situations, proposes a detailed neural circuit that satisfies these representational requirements, and demonstrates that the propagation of a suitable pattern of activity encoding an event can lead to the rapid and automatic formation of the requisite neural circuit within the HS. 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 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, several behavioral and functional imaging studies were designed based on predictions made by the SMRITI model. The results of this work were disseminated in articles and invited colloquia. Some of the results of the above work are described in [11, 12].

THE ROLE OF CORTICO-SUBCORTICAL LOOPS IN PLANNING AND WORKING MEMORY: A COMPUTATIONAL MODEL.

Clinical and experimental research over the last decade has implicated neuroanatomical 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 2002, Narayanan developed 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 [7].

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

[1] A. Ankolekar, M. Burstein,J. Hobbs,O. Lassila, D. Martin, S. McIllraith, S. Narayanan, M. Paolucci, T.Payne, K. Sycara, and H. Zeng. DAML-S: Web Service Description for the Semantic Web, International Semantic Web Conference (ISWC) Sardinia, June 2002.

[2] J. Bryant, N. Chang, R. Porzel, and K. Sanders. Where is natural language understanding? Toward context-dependent utterance interpretation. In Matthew Crocker, Frank Keller and Christoph Scheepers (eds). *Proceedings of the 7th Annual Conference on Architectures and Mechanisms for Language Processing* AMLaP, Saarbrücken, 2001.

[3] N. Chang, J. Feldman, R. Porzel and K. Sanders. Scaling Cognitive Linguistics: Formalisms for Language Understanding. In Malaka R., Porzel R. and Strube M. (eds) *Proceedings of the*

First International Workshop on Scalable Natural Language Understanding. European Media Laboratory. Heidelberg, 2002.

[4] N. Chang, S. Narayanan, and M. Petruck. Putting Frames in Perspective, The Nineteenth International Conference on Computational Linguistics (COLING 2002), August 24 -Sept. 1, Taipei, Taiwan, 2002.

[5] N. Chang, S. Narayanan, and M. Petruck. From Frames to Inference, Scalable Natural Language Understanding (SCANALU), Heidelberg, May 22-25, 2002.

[6] M. Garagnani, L. Shastri, and C. Wendelken. A connectionist model of planning via back-chaining search. In *Proceedings of the Cognitive Science Society Conference 2002*. pp. 345-350, Fairfax, VA, August, 2002.

[7] S. Narayanan. The Role of Cortico-Thalamic-Basal Ganglia Loops in Working Memory and Executive Control. *Computational Neuroscience (CNS* 2002). To appear in *Neurocomputing* 2003).

[8] S. Narayanan, C. Fillmore, C. Baker, and M. Petruck. FrameNet meets the Semantic Web: A DAML+OIL representation of FrameNet. Language Resources for the Semantic Web, AAAI 2002, Edmonton.

[9] S. Narayanan and S. McIlraith. Simulation, Verification and Automated Composition of Web Services. Eleventh International World Wide Web Conference (WWW2002), Honolulu, May 7-10, 2002. Selected to Appear in *Computer Networks*, 2003.

[10] R. Porzel and I. Gurevych. Towards Context-adaptive Utterance Interpretation. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistic, Philadelphia. 2002.

[11] R. Porzel and T. Marcinak. Encoding Spatial Localizations with Construction Grammar. In J. Ostman and J. Leino (eds) In *Proceedings of the Second International Conference on Construction Grammar*, University of Helsinki. 2002.

[12] L. Shastri. Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences*, 6 (4): 162-168. 2002.

[13] L. Shastri. A computationally efficient abstraction of long-term potentiation. *Neurocomputing*.44-46: 33-41. 2002.

[14] L. Shastri and C. Wendelken. Learning structured representations. *Neurocomputing*. In press.

[15] C. Wendelken.. The role of mid-dorsolateral prefrontal cortex in working memory: a connectionist model. *Neurocomputing*, 44-46: 1009-1016. 2002.

[16] C. Wendelken and L. Shastri. Combining belief and utility in a structured connectionist agent architecture. In *Proceedings of the Cognitive Science Society Conference 2002*. pp. 926-931, Fairfax, VA, August, 2002.

SPEECH PROCESSING

During 2002, Speech Group activities included a wide range of research topics, from work on improved front-end processing of the speech signal to back-end analysis of structure in speech from meetings. Along a different dimension, projects ranged from highly speculative explorations to involvement in actual system prototyping. By mid-year the group was deeply involved in DARPA's new speech recognition program called EARS. Launching this project comprised a major fraction of our efforts in 2002.

The year's Speech efforts were headed by senior research staff members Jane Edwards, Hynek Hermansky (OGI and ICSI), Nelson Morgan, Barbara Peskin, Elizabeth Shriberg (ICSI and SRI), Andreas Stolcke (ICSI and SRI), and Chuck Wooters. Additionally, Qifeng Zhu, formerly of Nuance Communications, joined us in October. Dan Ellis continued his work with us while primarily being a faculty member at Columbia University. George Doddington is also a welcome addition, consulting 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 other visitors. We continued to benefit greatly from joint work with our domestic and international colleagues.

The sections below describe a number of the year's major activities in speech processing. This year we have organized the report in terms of major projects: EARS (large vocabulary speech recognition and related tasks); the Meeting Recorder project, which is in the process of releasing a significant new corpus to 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. We also started some new work in speaker recognition, and continued earlier work on matching speech algorithms to somewhat specialized computer architectures. 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.

EFFECTIVE AFFORDABLE REUSABLE SPEECH-TO-TEXT (EARS)

Beginning in May 2002, the Speech group at ICSI began work on the DARPA "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 a lead site for the Novel Approaches project (along with team members SRI, University of Washington, OGI, Columbia, and IDIAP in Switzerland). 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, and LDC will be providing data. NIST is under contract to handle the evaluation process. For the Novel Approaches endeavor, the only team outside of the ICSI-led group is a project at Microsoft. The following sections describe the goals and a broad description of the directions for these two projects.

RICH TRANSCRIPTION

We are working to generate readable transcriptions of conversational and broadcast speech in multiple languages. "Readable" here means incorporating capitalization, punctuation, speaker markers, 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.

Current speech-to-text technologies still suffer from several key limitations that make them impractical for many potential applications, in both Government and commerce:

- Word-level transcription accuracy for spontaneous, conversational speech is only about 70%, and significantly lower in adverse conditions, such as in noisy environments or when the style of speech or topic of discussion are not well covered in the training data.

- Recognition output is an unstructured string of words, without proper capitalization, punctuation, paragraph breaks, speaker labels, and other common markings that make text readable and that would be expected from a good human transcriber. Furthermore, important information conveyed by speaking style, rather than the words themselves, is lost, such as the speaker's emphasis or any emotional qualities apparent in the manner of speaking.

- Large amounts of data and expensive hand-transcriptions and annotations are required to achieve state-of-the-art performance. For languages other than English, such data may be hard to obtain, and the linguistic expertise for data annotation and system development may not be available, further limiting performance of speech-to-text systems in those languages.

The core of the EARS Rich Transcription program is directed at overcoming these limitations, and seeks to

- Reduce the word error rate of automatic speech transcription by incorporating knowledge sources not currently captured in recognition systems, for improved accuracy and efficiency.

- Enrich the recognition output with multiple levels of metadata annotation, including sentence boundaries, types of utterances (e.g., question versus statement), changes in topics, and speaker identity. Also, disfluent speech might be edited, so that the intended meaning can be inferred.

- Combine word-level transcription and metadata into a unified recognition model, so that information about one can help improve the other. This will yield not only overall higher accuracy but also mutually consistent outputs.

- Develop techniques to quickly adapt the recognition system to new languages, speaking styles, and domains of discourse, without requiring full retraining on vast amounts of data as currently required.

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 incorporate innovations from our Novel Approaches effort into the Rich Transcription system. Under this heading we have also continued our work on multi-band systems, using both multi-layer perceptrons and graphical models to integrate spectrally local information into recognition. ICSI work in this area was primarily done by UCB graduate student Barry Chen with some guidance from Nelson Morgan.

In the area of *Portability,* ICSI Research Associate Yan Huang (with assistance from Chuck Wooters, Andreas Stolcke, as well as Jing Zheng of SRI) has taken the lead in developing a Mandarin Broadcast News (BN) recognition system based in SRI's existing technology. She was able to bootstrap acoustic models using a mapping of English to Mandarin phones, and achieved accuracies comparable to those reported by other sites for similar systems. We are now in the process of incorporating more advanced features of our English system, and have also started research specific to Mandarin. This includes the use of tone-specific phones, and the development of a highly effective iterative word segmentation algorithm that allows us to tokenize the non-whitespace-separated Mandarin character streams to facilitate language model training. From an initial baseline of almost 30% character error rate (CER) on the BN 1997 devtest set, we have reduced the CER by nearly 15% absolute (roughly half). As another part of the Portability task, we also began looking at the design of language-independent speech units, focusing on a self-organization method adapted from speaker segmentation (see below). Chuck Wooters is the primary ICSI researcher on this latter study, working with Swiss visitors Jitendra Ajmera and Micha Hersch.

Much EARS activity at ICSI has been under the *Metadata Extraction (MDE)* task, with ICSI scientist Barbara Peskin heading up the overall task. Elizabeth Shriberg, who is affiliated with both SRI and ICSI, also plays a leading role. The MDE activity has followed three main lines: work on "structural" metadata, including automatic detection of disfluencies and slash units (a pragmatically defined element corresponding to sentence-like units); speaker segmentation and labeling; and infrastructure development, working with the EARS community to formulate task definitions as well as annotation and evaluation specifications for this very new research direction.

Yang Liu at ICSI (in collaboration with Shriberg and Stolcke) has been working on a baseline MDE system built from SRI's "hidden event" detector [23], assessing performance on the slash unit and disfluency tasks, and exploring the impact of recognition errors and the differences between the Broadcast News and Conversational Telephone Speech (e.g., Switchboard) domains. This baseline system has now been enhanced with a repetition-pattern based language model for detection of repetition disfluencies and augmented with additional prosodic indicators for better detection.

We also began work on extending a new approach to speaker segmentation developed by our partner laboratory IDIAP [3], based on a Bayesian Information Criterion that does not require a "twiddle factor" to compensate for the number of parameters in determining the

optimum number of clusters for the segmentation. This was primarily a contribution from Swiss visitor Micha Hersch. One of the principals from the original IDIAP effort, Jitendra Ajmera, is now a visitor at ICSI.

NOVEL APPROACHES

We are studying replacements of the standard spectral envelope from a framewise stepped 25 ms analysis window (typically with cepstral transformation) as the speech representation of choice. This standard source of "front-end" features is inherently sensitive to a wide range of sources of variability that are unrelated to the linguistic units that we wish to recognize. However, radical changes in the front end are unlikely to bring major improvements without a concurrent redesign of the statistical models that underlie the recognition process. The ICSI component of the Novel Approaches program (which has the internal name of "Pushing the Envelope … Aside") will consist of both front-end and modeling research, but our early efforts have focused almost exclusively on front-end design. In this work, our approaches have benefited significantly from the methods developed under the Aurora project described in last year's activity report.

The primary components of the ICSI-led Novel Approaches effort are:

- Develop alternatives to the spectral envelope
- Incorporate multiple front ends across time and frequency
- Modify the statistical models to accommodate new front ends
- Design optimal combination schemes for multiple models

The first 6 months of this project were dominated by infrastructure: hiring (in particular, the recruiting of Qifeng Zhu), data set determination (both training and test), and the building of baseline recognition systems. For the latter, we decided to initially test on a limited vocabulary task (recognition of natural numbers), while training on a much broader data set including tens of hours of conversational speech. The goal was to avoid specializing our training on a limited test set. More recently we have begun exploring the use of a medium-sized vocabulary taken from the most frequent words in Switchboard.

ICSI is acting as the central site in this collaboration, and has the primary responsibility for the integration of novel approaches from the other sites into systems that can demonstrate performance improvements. In addition to this integration, ICSI consultant George Doddington is working to define evaluation materials, metrics, and diagnostics to guide progress. Finally, ICSI continues to develop novel components for speech recognition systems, both independently and in collaboration with the other sites.

In one set of experiments, we used the initial data splits that gave us small (.6 hour) Numbers data sets for development and testing. Having previously established baselines, with these data, we sought to explore using two candidate approaches developed in OGI-ICSI-Columbia collaborations: Tandem processing (using a neural network for discriminative transformation of features such as PLP) and TRAPs (incorporating long [250 ms – 1s] temporal trajectories within small [1-3 critical bands] subbands). In addition, we incorporate the baseline features (PLP+energy with 1st and 2nd derivatives, plus speaker

normalization). The main point of these experiments was to refine combination approaches, such as combination via Karhunen-Loeve Transformation (KLT), adding log probabilities, adding weighted probabilities (with weights determined from inverse stream entropy), or by simple feature concatenation. Ultimately, the best methods for these experiments appeared to be simple log probability addition and KLT combination. The entropy approaches appear to be extremely effective for additive noise, but they may not be the best for clean speech; however, we will continue this exploration as we move to conversational speech. So far, though, our best results give us an improvement from 3.8% to 3.1% word error rate on about 5000 words, using either KLT combination or log probability addition. A key issue in this combination was dimensionality reduction via KLT. While we got many specific results for each case, we don't know in general how to fix the number of output features, as this may well depend on the size and style of test sets. However, we now have procedures for determining these values given a development set, and our experience tells us that this step may be crucial in optimizing performance.

We are also working on alternative methods at the level of neural networks for processing TRAP-style information to recognize phones in TIMIT. So far the best approach appears to be to separately train nets on each temporal pattern to discriminate between phones, then "behead" each net by taking the hidden layer activations and use them to train a combination network. The phone error rate is 32.7% when the full networks are used, while the combined "beheaded" nets yield an error rate of 29.3%.

While this work was being done, we also designed a larger Numbers data split for Novel Approaches development. We defined three sets: one tuning set (4165 words, 0.566 hours), one dev set (9430 words, 1.28 hours), and one test set (9708 words, 1.33 hours). We re-ran our baseline tests and got similar results. We are now in the process of repeating the most promising experiments from the Tandem/TRAP combination work described above using the larger sets to see if the conclusions still hold. We are of course sharing all of these data sets (as well as the specification of the training set) with the other sites in our team, but in addition we are sharing them with the Microsoft Novel Approaches team so that we can have some points of comparison.

Finally, a critical ICSI activity is work with the other sites on the creation of a domain transition plan for transitioning the Novel Approaches research activities from the current Numbers domain to the target domain of (unrestricted) conversational speech. This is important both for deployment and for research collaboration and guidance. To this end a study was made of restricting speech input to utterances containing only the most common words, using the ISIP transcriptions of the Switchboard-I corpus. George Doddington (an ICSI consulting scientist) found that 17% of the word tokens (relative to the entire conversational telephone speech portion of the dry run corpus) were covered by utterances that contain only the 500 most frequently occurring words. The motivation was that by selectively using utterances composed only of these most frequent words (e.g., 500), we could switch from read to conversational speech while still limiting the problem size (and decoding difficulty) to something that was manageable when working with novel approaches. One drawback to this plan, however, is that such utterances tend to be "weird"; that is, somewhat atypical and low in semantic content. However, by expanding the chosen utterance list to those containing at most 10% Out-Of-Vocabulary (OOV) words, we could get a somewhat more representative set (covering about half of the utterances in Switchboard) while still keeping the overall OOV rate down to manageable level (about

3.2%). Our training data, however, will continue to incorporate utterances with the unlimited Hub 5 vocabulary.

THE MEETING RECORDER PROJECT

In the 2001 Activity Report we described an ongoing ICSI project on processing speech from meetings (in collaboration with other sites, principally SRI, Columbia, and UW). Here we report our more recent progress in this effort, including: the completion for release of a 75-meeting corpus (over 70 meeting-hours and up to 16 channels for each meeting); the development of a prosodic database for a large subset of these meetings, and its subsequent use for punctuation and disfluency detection; the development of a dialog annotation scheme and its implementation for a large subset of the meetings; and the improvement of both near-mic and far-mic speech recognition results for meeting speech test sets. This work was initially done under DARPA sponsorship, but by the end of 2002 support had shifted to a combination of European Union funding (under the Sheffield-led M4 project) and Swiss funding (as part of the Swiss National Science Foundation-funded IM2 research network, led by our partner laboratory IDIAP). In addition, we continued to participate in an NSF-ITR program called "Mapping Meetings" in collaboration with the University of Washington, Columbia University, and SRI.

CORPUS COLLECTION

We have concluded the collection and preparation for distribution of the ICSI Meeting Corpus, which contains audio and transcripts of natural meetings recorded simultaneously with head-worn and table-top microphones. The corpus contains 75 meetings of 4 main types and 53 unique speakers. We will deliver the corpus to the LDC by April 2003, and expect it to be available through the LDC by the fall of 2003. Details on the corpus can be found in [13].

We have published many papers related to the corpus, including research on automatic transcription, speech activity detection for segmentation, overlap analysis, applications, prosody, automatic punctuation, noise robustness, and reverberation. For an overview, see [17], which appears in the special session on Smart Meeting Rooms in the ICASSP 2003 Proceedings. For a complete listing of publications from ICSI on the Meeting Corpus, see http://www.icsi.berkeley.edu/Speech/mr/index.html .

In addition to the data released with the corpus, we also continue to annotate the corpus with additional information, including dialog act labeling and prosodic features. We hope that others will also contribute to the corpus, either with additional meeting data, or with more annotations of the existing data.

SENTENCE SEGMENTATION AND DISFLUENCY DETECTION

In addition to fundamental work on automatic word-level transcription, the Meeting Corpus supports a variety of "Rich Transcription" tasks. In this section, we report on efforts to automatically detect sentence boundaries and disfluencies by means of prosodic information and lexical cues. Since recognition error rates are still quite high for this domain, prosodic cues may take on greater importance than they would for a domain with low word error rates.

We processed and analyzed data from three main meeting types, with between 3 and 8 speakers each. There was a total of 31.9 hours of speech, where this duration excludes long silence regions but counts overlapped speech multiple times. There were 306,957 transcribed words in this sub-corpus. The transcriptions and turn-level segmentations were revised and supplemented with various annotations including markings for disfluencies and incomplete sentences. For segmentation, recognition, and forced alignment to reference transcripts (useful for a baseline for event detection), the close-talking microphone signals were used. The speech recognition system used for these experiments was a simpler system than what we have used for ASR evaluation purposes, and achieved word error rates of roughly 45% on native speakers and 72% on non-natives.

The task for this study was to detect interword "events" corresponding to a sentence boundary, or to a disfluency or incomplete sentence breakpoint. The event classifier made use of features extracted in four main categories: pause and duration, fundamental frequency, energy, and context. Pause durations were computed based on alignments, and were fairly robust to recognition errors. Phone durations were obtained from ASR or forced alignments, and were normalized by Switchboard phone durations. Pitch features were derived from the ESPS pitch tracker get_f0, followed by median filtering and a piecewise linear fit. Further normalization was done using baseline F0 values determined by a log-normal tied mixture model [26]. Energy features were also computed, and were normalized by channel statistics. Non-prosodic contextual features included speaker gender, native or non-native, and whether or not the speech included some overlap of multiple speakers. All of these features were modeled by decision trees that produced posterior probabilities for the various event types. In addition, trigram language models (LMs) predicted the same events based on word context, and a combined model integrated prosodic and lexical features, using the hidden Markov model approach previously developed for hidden event modeling [23].

In our test set, 9% of word boundaries were sentence breaks, 10% were disfluencies or incomplete sentence breaks, and the remaining 81% were fluent boundaries. Similar numbers characterized the training set. Consequently, "chance" classification accuracy for this 3-way classification task is 81%, which could be achieved by simply calling all boundaries fluent.

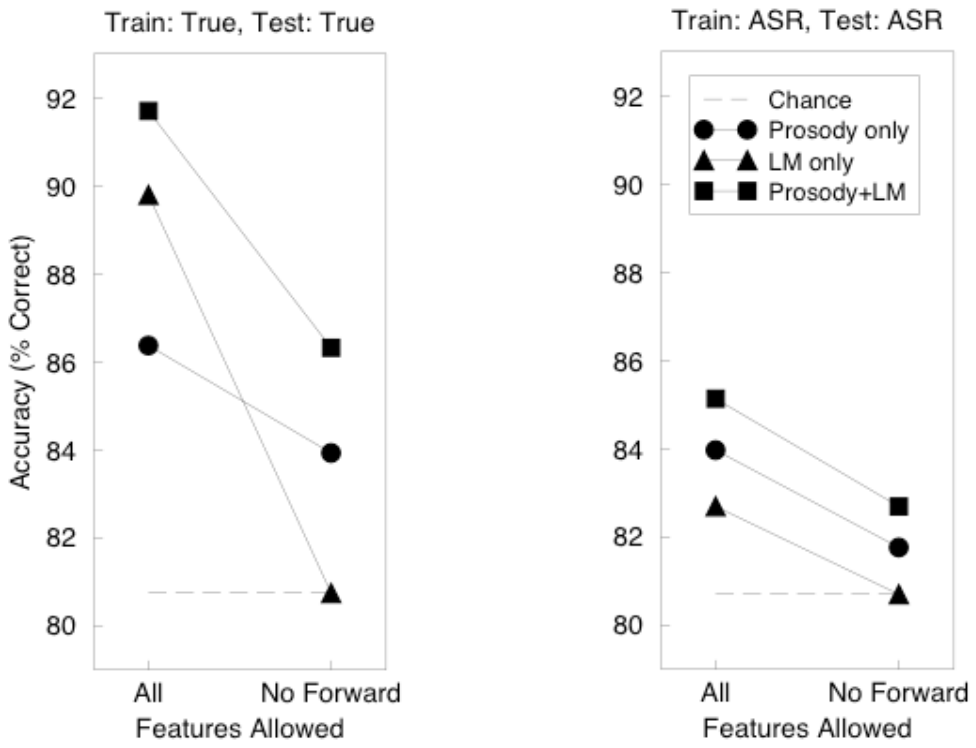

Figure 1: Event detection accuracy (in \%) using different models and different train/test conditions. ``True'' = true words (forced alignment);``ASR'' = 1-best recognizer output; ``LM'' = language model.

Figure 1 shows event detection accuracies for prosody-only, LM-only, and combined models under four conditions: with reference versus automatic word transcripts, and with/without features that look "forward" in time. The latter condition is of interest for future systems that aim to process meetings online, e.g., to give real-time assistance to meeting participants.

As shown, all classifiers (except the LM without forward-looking features) perform significantly above chance, with the combined model achieving 92% accuracy with correct words, and 85% with ASR output. The combined model is uniformly better than either prosody or LM alone. As one might expect, the prosodic model alone degrades less as a result of word recognition errors than the LM, but is also more robust to the loss of forward-looking features. Inspection of the prosodic model shows that it relies primarily on duration features, corresponding to vowel lengthening and pause durations. More details and additional results can be found in [7].

DIALOG ACTS: ANNOTATION AND OBSERVATIONS

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 [14] 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.

As of the end of 2002 about 10 hours of meetings have been fully annotated with dialog act labels, and more are in process. For this subset, about 65% of the utterances are statements, 9% are questions, and 26% are backchannel remarks. About 14% of all utterances are disrupted (for instance by another speaker's interjection), and about 7% of the utterances end on a rising fundamental frequency. Some other labels that occur over 1% of the time include joke, self-repeat, completing someone else's utterance, or repeating someone else's utterance. This work is still quite preliminary, but we already see some interesting trends. For instance, it is "common knowledge" that questions in English are characterized by rising pitch. However, in the data we have seen so far, the classification is not so simple. Of the utterances which end in rising intonation, 60% are questions, but a hefty 33% are statements. (The remaining 7% are a mix of backchannels, floor-grabbers, and other occasional types.) So rising F0 alone is not a reliable predictor of questions.

RECOGNITION OF NATURAL MULTI-PARTY SPEECH

In 2002, meeting data formed a track of NIST's Rich Transcription evaluation, RT-02, and we report here on experiments associated with that task, using meeting data provided by ICSI and by other meeting collectors: NIST, LDC, and CMU.

To establish a baseline for automatic meeting transcription, we first ran our Switchboard-trained recognition system, based on SRI's Decipher technology, on the meeting data. While there was a small cost for downsampling the speech data to accommodate the telephone bandwidth of Switchboard, we felt that the conversational speech of Switchboard offered the best match to the natural speaking style found in meetings. Without any meeting data used to train the recognizer, we obtained an average word error rate (WER) of 36.0% on the collection of two 10-minute meeting excerpts from each of the four sources (with WERs ranging from 25.9% to 47.9% for the various sources), using the close-talking channels. This result is surprisingly close to that obtained in the standard conversational telephone speech tasks (such as English Hub 5) and indicates that meeting data -- at least as recorded via close-talking microphones -- is an accessible task using current technology.

Next we wished to explore the cost incurred by not having language model training data geared to the Meeting task. For this experiment we used transcripts from the 28 ICSI meetings at that time transcribed (excluding the four -- two development and two evaluation

-- that we contributed to NIST's evaluation). This amounted to approximately 270k words of transcribed speech, including 1200 new words not already in the recognizer's vocabulary. We tested the new language model only on the two ICSI evaluation sets, again using the close-talking mics but a somewhat simplified recognition protocol. Word error rates dropped from 30.6% using the original Switchboard LM to 28.4% using an interpolated language model that combined the small Meeting LM with the Switchboard LM (with interpolation weights estimated from the two ICSI development meetings). With the addition of the new meeting data, the out-of-vocabulary rate dropped from 1.5% using the Switchboard LM to 0.5% using the interpolated Meeting LM.

We also explored the impact of using far-field, rather than close-talking, recordings. We found that word error rates for the tabletop mics were virtually double those for the close-talking ones: error rates rose to an average of 61.6%, with scores from the various sources ranging from 53.6% to 69.7%. We found that performance on far-field mics could be improved by applying Wiener filtering techniques we developed for the Aurora program [2]. On our ICSI/LDC/CMU development set, this resulted in a decrease from 64.1% WER to 61.7%. Still, the error rates are so much higher than for the near mic that we found it best to conduct our research using read digit strings that were collected at ICSI using the same meeting room, talkers, and microphones as for the conversational speech. (The data from this read-speech "digits task" is being released as part of ICSI's official Meeting Corpus.) This work is described later in this report as part of our general effort on acoustic robustness for speech recognition.

EXPLORING MEETING STRUCTURE

Jane Edwards has been working on the Meeting Maps project, with the help of her student assistant, Jennifer Alexander. Her goal is to identify linguistic and prosodic correlates of meeting structure which can be used by colleagues in automatically generating maps of content in meetings. She is particularly interested in "discourse markers." These are items, such as "so", "but", "nevertheless" and many others, which are used by speakers to explicitly indicate connections they perceive between propositions in the discourse (e.g., [20]). In some cases, a particular lexeme may have some uses which are not connective in nature. For example, the "so" in "that's so easy" is not a discourse connective, whereas it is one in "So, that's easy". Often the two uses are prosodically distinct, even if that is not always the case [21]. Related work includes [12].

To explore the use of discourse markers to cue topic boundaries or continuers, we are building on the work of colleagues and extending it. To initialize a database of meetings segmented by topic, we hand segmented some meetings into topics and compared our segmentations with those done at some of our partner sites. They are sufficiently similar that we will be able to use their segmentations to broaden our database. For the prosodic features, we will be using the same rich prosodic database used in the sentence segmentation and disfluency detection work described above.

This study is still at a preliminary stage, but main questions will be:

- Which discourse markers are most relevant to generating content maps?
- Which prosodic cues are most useful in discriminating connective from non-connective uses?

OTHER EFFORTS

SPOKEN DOCUMENT RETRIEVAL AND TOPIC VISUALIZATION: In work jointly supported by the Meeting Maps project and the EU project M4, ICSI visitor and long-time associate Steve Renals ported his spoken document retrieval system, developed for Broadcast News, to the Meetings domain. Rather than attempting an a priori topic segmentation of the meetings, Renals simply used sequences of overlapping 30 sec excerpts as his retrieval units, with Porter stemming of words but no stoplist. The query-document match employed a term frequency and inverse document frequency (tf.idf) measure. Using a simple set of 10 queries (which included some keywords which were out-of-vocabulary for the ASR engine), he found surprisingly little degradation between performance on hand transcripts and on ASR output, when evaluated using standard precision-at-n-documents measures. Renals also experimented with techniques for visualizing topic structure within a meeting by generating a self-similarity matrix where cell (i,j) represented a tf.idf-based similarity measure of the relatedness of minute i of the meeting to minute j. The largely block-diagonal structure aligned nicely to a human-generated segmentation of topic content and provided an intriguing view of the topic structure. Details can be found in [18].

SPEAKER TURN PATTERNS AND "TALKATIVITY": Another Meeting Maps and M4 project involved the segmentation of meetings not by topic but by speaker turn patterns. This work, performed by ICSI associate Dan Ellis, introduced a speaker transition matrix for each minute of the meeting, representing the probability that speaker i follows speaker j as the dominant speaker. He then used the Bayesian Information Criterion to segment the sequence of such matrices into chunks, seeking change-points in the pattern of speaker turn-taking. This provided a very different view of the meeting structure than that generated by topic. Ellis also introduced a measure of speaker "talkativity" to calibrate individual meetings based on the degree of participation of the attendees, as normalized by their baseline propensity to speak. Again, see [18] for further details.

SUMMARIZATION: Renals also explored the application of extractive text summarization methods to the meetings data. This work was based in an information retrieval framework. The set of indexed meetings was queried, resulting in a set of meeting segments that were (potentially) relevant to the query. Each segment was summarized using a technique known as maximum marginal relevance (MMR). MMR constructs a summary out of extracts of the returned segments, balancing relevance to the query against diversity with respect to already extracted parts.

HOT-SPOTTING: For an investigation of the underlying interactive structure of conversations, ICSI visitor Britta Wrede hand-annotated 32 meetings for "hotspots", i.e. places where participants get especially involved. Hotspots should be related to the information structure of meetings and thus be useful for information retrieval or summarization tasks. On a more detailed level, individual turns of these hotspots were labeled with regard to their involvement. Results of a preliminary experiment suggest that there is high inter-labeler agreement on such turns, indicating that "involvement" is a highly

salient perceptual category. Another experiment is planned which focuses on the perception of whole hotspots. In further work the acoustical characteristics of involved turns will be analyzed and used to train a classifier. Automatic detection of hotspots based upon this classifier and further information from the labeling of hotspots is the final goal of this project.

ACOUSTICALLY ROBUST SPEECH RECOGNITION

As noted above, the EARS Novel Approaches project provides a major focus for the development of new acoustic representations that are more likely to be stable across a range of conditions of speaking style and acoustic environment. For EARS, since the emphasis is on telephone conversations, the primary factor causing degradation is the conversational speech itself. However, there are many practical applications (including the Meetings application using far-field mics) in which the specific acoustic factors of noise and reverberation are the prime culprits in causing speech recognition errors. In previous years, our focus task for this area of research was the Aurora standardization process organized by the European Telecommunications Standards Institute. Some of this work continued in early 2002, but by later in the year we were largely working with speech from distant (3-6 feet) microphones. This was intended to model the situation in which speech from a number of people was processed by a portable device, such as one might use to record a conversation in a meeting (with mutual consent, of course). Speech from such a distance generally has significant components from environmental noise and delayed signal components due to room reverberation. The most convenient corpus for this work was the set of digits strings recorded just before or after our recorded meetings, as described previously. Working with the digit strings data, we attempted to deal with far-field acoustics using the model of a convolutional distortion (the room response) followed by an additive distortion (background noise) [10].

For the additive distortion, we used the previously mentioned noise reduction algorithm based on Wiener filtering (with typical engineering modifications such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor).

For the convolutional distortion, we applied the technique of long-term log spectral mean subtraction, which uses a similar principle to cepstral mean subtraction to reduce convolutional effects but with an FFT window length longer (e.g., 1 s) than the typical 20-30 ms used in cepstral analysis for ASR. In later experiments we found that under some conditions (such as if we trimmed off the smallest half of the values used in computing the mean for each spectral bin) we could have good results on our meeting room data with shorter windows as well. In all cases, the error rate reduction corresponded to a model of reverberant speech as being close-talking speech convolved with an unknown impulse response (and thus producing an additive component in the log spectral domain).

The noise reduction and the log spectral subtraction each individually improved performance on the digits task, with the log spectral subtraction having the greater effect, and had a cumulative effect when used together. However, even when using both, we still observed a huge difference between near-mic and far-mic cases, obtaining 2.7% and 7.2% word error rates respectively when we applied the techniques as pre-processing for the Aurora recognizer described in [11], which was trained on TIDIGITS digit strings data.

While the far-mic case described above was implemented using a high-quality (PZM) microphone, we have also been interested in the level of ASR performance that could be obtained using inexpensive electret microphones such as might be found in a PDA. For this case, we found that the effect of using Wiener filtering and log spectral subtraction was even greater. Without such processing, error rates were significantly worse than for the better microphone, but error rates were roughly comparable to the PZM scores after the processing.

The techniques that we have emphasized have typically been inspired by the human example, though in retrospect each of them makes sense on the basis of signal processing considerations alone. However, they have not typically mimicked human physiology, but rather function from a psychophysical standpoint – in other words, implementing the observed behavior rather than the biological function per se. However, we recently have begun work with Infineon researcher Werner Hemmert to explicitly use models of early auditory function in order to make ASR more acoustically robust. To simplify the early experiments, we went back to the Aurora noisy digits task, and integrated a detailed cochlear model with one of our standard systems. This has presented numerous challenges. For one thing, the cochlear model has much more time and frequency resolution than we believe to be important for monaural ASR – the resolution might be critical for pitch detection and spatial localization in the human system, but these are not our tasks at the moment. It is straightforward to simplify the cochlear model to better match the typical feature resolution for our ASR systems, but it is likely that the potential advantages are lost in this case. More detailed models are also more difficult to integrate with currently workable statistical models (although future ones being developed under EARS may be a better match). This work has really just begun, and we hope to report further on progress in next year's report.

In cooperation with visiting scientist Michael Kleinschmidt of the University of Oldenburg, Germany, we investigated a new approach to the front end of automatic speech recognition systems, using a set of two-dimensional Gabor filters with varying extents in time and frequency and varying ripple rates to analyze a spectrogram. 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. Good results were obtained in a noisy digit recognition task. We presented this work at ICSLP 2002 in [16].

SPOKEN LANGUAGE SYSTEMS

SmartKom is a project funded by the German Ministry of Education and Research, aiming to develop a multi-modal dialog system that can assist humans by interacting with speech, gesture, and facial expression (see http://www.smartkom.com). ICSI is one of about a dozen academic and industrial partners on the project and is responsible for developing the English-language speech recognizer for the system, as well as porting various other language-dependent components from their original German version to English.

NOISE AND ACOUSTICS COMPENSATION

As noted in the previous section, ICSI continued its research in the area of compensation algorithms for both ambient noise and room acoustics (reverberation). Much of this work has been supported by the SmartKom program. The most fully developed components of this work (in particular, the Wiener-style noise reduction) have been implemented in our SmartKom system.

LANGUAGE MODEL INTEGRATION

In past versions of the English SmartKom system, the recognizer used class-based language models for better generalization and to incorporate domain knowledge. However, the class-based LM was statically compiled into an equivalent word-based N-gram LM since the ICSI decoder was not able to handle class-based LMs directly. In order to support class-based models more effectively, including the ability to update class definitions in the running recognizer (as required by current SmartKom systems), we have implemented a tight software integration of the decoder with the SRILM toolkit, which supports class-based and many other advanced types of LMs. The tight integration allows the decoder to use any LM supported by the toolkit (subject to the Markov constraints of the recognizer), and to dynamically update the models when triggered by an external signal. We have also enabled the decoder, which is at the core of the English recognition module in SmartKom, to dynamically modify its language model and to update word class definitions. While this functionality was not yet needed for the current demonstration system, it will be essential for the final SmartKom system. Furthermore, since both ICSI's decoder and SRILM are freely available for research, we will be able to make the results of our integration work available to the research community at large.

SYSTEM DEVELOPMENT AND DEMONSTRATION

ICSI played a major role in creating the English-language version of the SmartKom Mobile demonstration system (SmartKom 3.2). Our role included the translation of the language analysis and text generation databases, the creation of updated language model and pronunciation lexicon for speech recognition, the porting of the speech synthesis to English in cooperation with IMS, and system integration and testing in cooperation with DFKI.

The system – a personal travel information assistant – was then demonstrated during a 3-day period at the International Conference for Spoken Language Systems (ICSLP-2002) in Denver, Colorado. We had very good audience interest, and received comments to the effect that the system was impressive.

The hardware configuration used for the demonstration system was assembled and tested within just a few weeks prior to the conference. It consisted of 2 IBM ThinkPad laptops, one running Windows with 512MB of memory, the other running Linux with 640MB, plus a Compaq iPaq running Pocket Linux for display output and pointer input, all connected by a Linksys switch and wireless access point. System responsiveness was acceptable using this hardware configuration. We placed the testbed GUI on the Windows system's screen to make better use of the available screen real estate and facilitate explanation of the system to the audience. The demo dialogs and even variations thereof (e.g., using pointing gestures for pedestrian navigation requests) worked surprisingly well, given the short time spent on development and testing.

SPEAKER RECOGNITION

ICSI scientist Barbara Peskin participated in one of the working groups at the Johns Hopkins 2002 Summer Workshop (hosted by the Center for Language and Speech Processing, and sponsored in part by NSF), exploring the use of higher-level features in speaker recognition systems. Most speaker recognition systems today use only low-level acoustic features, decomposing the speech stream into a series of 10-20 msec frames using

standard signal processing techniques, and modeling these frames as essentially independent events via Gaussian mixture models (GMMs) for "generic speech". Such models ignore many sources of information – word usage, prosodic characteristics, pronunciation patterns, idiosyncratic laughs or other nonspeech events – which may contain tremendous speaker-identifying power. The self-proclaimed "SuperSID" team (SID = Speaker ID) set as its summer's work the systematic study of these alternate knowledge sources. Armed with a large number of resources, including automatic speech recognition transcripts, automatic phone decodings, a database of prosodic and lexical features, and more, all based on a standard test set of conversational telephone speech, the group of researchers gathered together for six weeks of intensive exploration into the value of information sources at a variety of levels: frame-based acoustic, phone, prosodic, lexical, and conversational. The results were impressive: even with a baseline GMM system achieving equal error rates as low as 0.7% with 8-conversation-side training, the systems constructed at the workshop using alternate knowledge sources combined to match that performance using no frame-level GMM input at all, and when fused with the baseline GMM system were able to drive the equal error rate down to 0.2% [19].

MATCHING SPEECH ALGORITHMS TO COMPUTER ARCHITECTURES

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 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, the primary focus of our current effort is on this portion of the speech system.

Our initial implementation consists of a small vocabulary system. With a small vocabulary, it is not necessary to share states among similar words; rather, one can evaluate all the words separately. This allows an efficient, regular implementation. On IRAM, we arrange batches of words with total length equal to the vector length. On Imagine, we batch words such that the total length will fit in the cluster memory. We are in the process of analyzing the results of this approach.

Future work includes running a large vocabulary system on these architectures. This involves picking a search order that will maximize re-use of states from previous searches (e.g. if the word "architecture" has already been processed, most of the work can be re-used for the word "architectural"). Language modeling, beam pruning, and least-upper-bound path calculations may also be accelerated on these architectures.

REFERENCES

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI Features for ASR. ICSLP-2002, Denver, Colorado, September 2002.

[2] A. Adami, S. Kajarekar and H. Hermansky. A New Speaker Change Detection Method for Two-Speaker Segmentation. ICASSP-2002, Orlando, Florida, May 2002.

[3] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan. Unknown-Multiple Speaker Clustering using HMM. In *Proceedings. ICSLP-2002*, Denver, Colorado, September 2002.

[4] J.C. Ang. Prosodic Cues For Emotion Recognition In Communicator Dialogs. M.S. Thesis, University of California at Berkeley, December 2002.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog. ICSLP-2002, Denver, Colorado, September 2002.

[6] D. Baron. Prosody-Based Automatic Detection of Punctuation and Interruption Events in the ICSI Meeting Recorder Corpus. M.S. Thesis, University of California at Berkeley, May 2002.

[7] D. Baron, E. Shriberg, and A. Stolcke. Automatic Punctuation and Disfluency Detection in Multi-Party Meetings Using Prosodic and Lexical Cues. ICSLP-2002, Denver, Colorado, September 2002.

[8] S. Chang, A Syllabl. Articulatory-Feature, and Stress-Accent Model of Speech Recognition. PhD Dissertation, University of California at Berkeley, Sept 2002.

[9] D. Gelbart, Reducing the Effect of Room Acoustics on Human-Computer Interaction. Avios-2002, San Jose, California, May 2002.

[10] D. Gelbart and N. Morgan, Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition. ICSLP-2002, Denver, Colorado, September 2002.

[11] H. G. Hirsch and D. Pearce. The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. ISCA ITRW ASR2000, Paris, 2000.

[12] J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19, 501-530, 1993.

[13] Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. The ICSI Meeting Corpus. In *Proceedings of ICASSP-2003*. Hong Kong, April 2003.

[14] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard-DAMSL Labeling Project Coder's Manual, Tech. Rep. 97-02, University of Colorado, Institute of Cognitive Science, Boulder, Colorado, 1997, http://www.colorado.edu/ling/jurafsky/manual.august1.html

[15] M. Kleinschmidt. Spectro-temporal Gabor Features as a Front End for Automatic Speech Recognition. Forum Acusticum 2002, Seville, September 2002.

[16] M. Kleinschmidt and D. Gelbart. Improving Word Accuracy with Gabor Feature Extraction. ICSLP-2002, Denver, Colorado, September 2002.

[17] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations. In *Proceedings of ICASSP-2003*. Hong Kong, April 2003.

[18] S. Renals and D. Ellis. Audio Information Access from Meeting Rooms. In *Proceedings of ICASSP-2003*. Hong Kong, April 2003.

[19] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition. In *Proceedings of ICASSP-2003*. Hong Kong, April 2003.

[20] D. Schiffrin. *Discourse Markers*. Cambridge; New York: Cambridge University Press, 1987.

[21] L. Schourup. Tutorial overview: Discourse markers. *Lingua*, 1999, 107, pp. 227-265, 1999.

[22] E. Shriberg and A. Stolcke. Prosody Modeling for Automatic Speech Recognition and Understanding. In *Proceedings of the Workshop on Mathematical Foundations of Natural Language Modeling*, M. Ostendorf, S. Khudanpur, R. Rosenfeld (eds.), Institute for Mathematics and its Applications, Minneapolis.

[23] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, vol. 32, no. 1-2, 2000.

[24] S. Sivadas and H. Hermansky. Hierarchical Tandem Feature Extraction. ICASSP-2002, Orlando, Florida, May 2002.

[25] P. Somervuo, Speech Modeling Using Variational Bayesian Mixture of Gaussians. ICSLP-2002, Denver, Colorado, September 2002.

[26] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling Dynamic Prosodic Variation for Speaker Verification. ICSLP-98, Sydney, 1998.

[27] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg. Using Prosodic and Lexical Information for Speaker Identification. ICASSP-2002, Orlando, Florida, May 2002.

BERKELEY CENTER FOR THE INFORMATION SOCIETY

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; and (3) the digital opportunities program called "Berkeley Foundation for Opportunities in Information Technology", led by Orpheus Crutchfield.(4) A pilot project on IT in Ghana, led by G. Pascal Zachary.

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 the best doctoral students in this research area.

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.

During its first three months, starting in Fall 2002, the key events of the Center include:

- publication of the book Manuel Castells and Pekka Himanen, The Information Society and the Welfare State: The Finnish Model (Oxford University Press, 2002)
- two high-level seminars on September 26 and October 17
- a working seminar from October to November
- a public talk by Pekka Himanen to introduce the Center for the UCB campus on December 9.