INTERNATIONAL COMPUTER SCIENCE INSTITUTE ACTIVITY REPORT 2001

CONTACT INFORMATION

POSTAL ADDRESS	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
INTERNET	http://www.icsi.berkeley.edu

PRINCIPAL 2001 SPONSORS

AT&T
Intel
Qualcomm
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 2001 SPONSORS

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

CORPORATE OFFICERS

Prof. N. Morgan (President and Institute Director)
Prof. S. Shenker (Vice President)
Dr. J. Cohen (Secretary and Treasurer)

BOARD OF TRUSTEES

Dr. H. Ahmadi, AT&T Labs-Research
Prof. E. Berlekamp (Chair), UC Berkeley
Prof. H. Bourlard, IDIAP and EPFL, Switzerland
Vice Chancellor B. Burnside, UC Berkeley
Dr. J. Cohen, VoiceSignal Inc.
Dr. A. Goldberg, Neometron
Dr. G. Heinzinger, Qualcomm
Mr. C. Higgerson, Comventures
Prof. R. Karp, ICSI and UC Berkeley
Mr. P. Lizcano, Comision Interministerial de Ciencia y Tecnolologia
Prof. N. Morgan (Director)
Prof. C. Papadimitriou, CS Division Chair, UC Berkeley
Dr. I. Reitmaa, TEKES
Prof. D. Schütt, Siemens AG
Prof. W. Wahlster, DFKI GmbH

SENIOR TRUSTEES

Prof. G. Barth, Dresdner Bank AG
Prof. J. Feldman, ICSI and UC Berkeley
Prof. G. Goos, U. Karlsruhe
Prof. D. Hodges, UCB, past ICSI Chair
Mr. R. Kay, former acting Director of ICSI (until 1988)
Prof. H. Schwärtzel, TU Munich
Prof. N. Szyperski, U. of Cologne; former head of the GMD; Founder of ICSI.

2001 VISITORS IN SPONSORED INTERNATIONAL PROGRAMS

NAME	COUNTRY	GROUP	AFFILIATION *
Carmen Benitez	Spain	Speech	MCYT
Jose Colas	Spain	Speech	MYCT
Juan Peire	Spain	Networking	MYCT
Miguel Sanchez	Spain	Networking	MYCT
Mikko Harju	Finland	Speech	TEKES
Lauri Hetemaki	Finland	Haas	TEKES
Pasi Koistinen	Finland	Haas	TEKES
Jan-Ola Ostman	Finland	FrameNet	TEKES
Ernst Althaus	Germany	Algorithms	DAAD
Michael Bahr	Germany	Networking	Siemens
Hans Boas	Germany	AI/FrameNet	DAAD
Sven Buckholz	Germany	Networking	Daimler
Stefan Bleck	Germany	Haas	AIF
Tanja Falkowski	Germany	Haas	Tech. U. Braunschweig
Andreas Franz	Germany	Haas	CDTM
Christian Gill	Germany	Haas	AIF
Matthias Hecking	Germany	Networking	FGAN
Michael Kleinschmidt	Germany	Speech	U. Oldenburg
Birger Kollmeier	Germany	Speech	U. Oldenburg
Christian Kreibich	Germany	ACIRI	Tech. U. Munich
Jonathan Landgrebe	Germany	Haas	CDTM
Peter Pfannes	Germany	Haas	CDTM
Thilo Pfau	Germany	Speech	DAAD
Robert Porzel	Germany	AI	EML
Josef Ruppenhofer	Germany	AI	ICSI
Stephanie Schnicke	Germany	Haas	CDTM
Joachim Sokol	Germany	Networking	Siemens
Felix Treptow	Germany	Haas	CDTM
Dietmar Tutsch	Germany	Networking	DAAD
Matthias Westermann	Germany	Algorithms	DAAD
Wilbert de Graaf	The Netherlands	Networking	KPN

* CDTM: Center for Digital Technology and Management
DAAD: Deutscher Akademischer Austauschdienst
EML: European Media Lab
FGAN: Forschungsgesellschaft für Angewandte Naturwissenschaften e. V. http://www.fgan.de/FGAN/En/
MYCT: Ministerio de Ciencia y Tecnologia, Estado de Politica Cientifica y Tecnologia
TEKES: Finnish National Technology Agency

Contents

INSTITUTE OVERVIEW – MARCH 2002

The International Computer Science Institute (ICSI) is an independent, nonprofit basic research institute affiliated with the University of California campus in Berkeley, California. Its establishment was motivated by a 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.

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 – 2001 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.

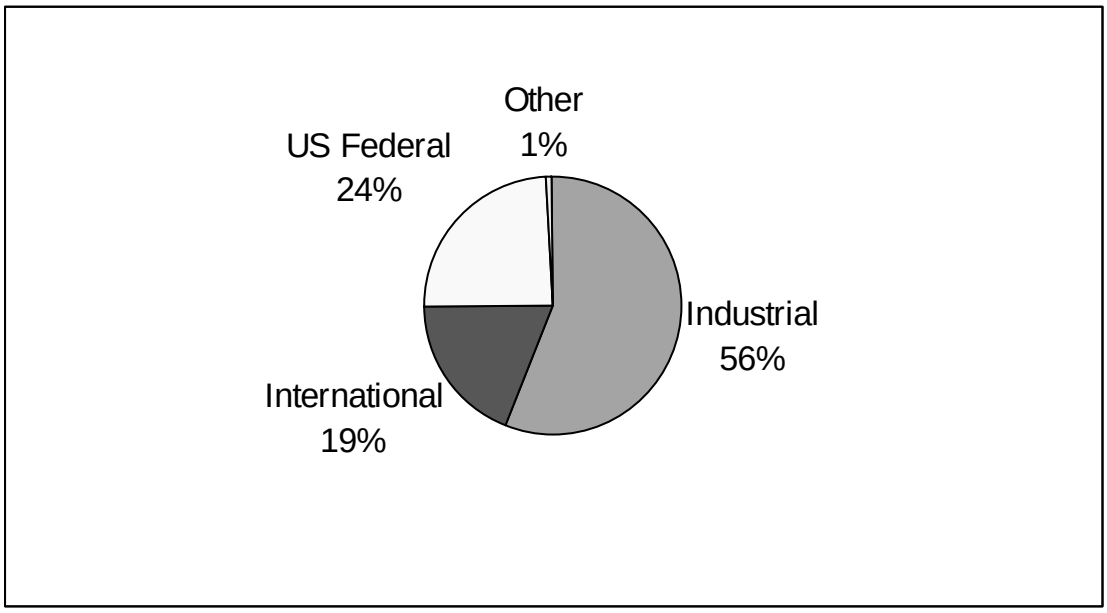


Figure 1: Distribution of sources of ICSI revenue for 2001.

US Federal funding comes from a range of grants to support research Institute-wide. Most of this funding comes from the National Science Foundation and DARPA. International support in 2001 came from government and industrial programs in Germany, the Ministry of Science and Technology in Spain, and the National Technology Association of Finland, with additional support from the European Media Lab and the European Union. Industrial support in 2001 was primarily provided by AT&T, Intel, Nortel, and Qualcomm with additional sponsorship from IBM, and Xerox.

Revenues increased significantly in 2001 (about 18% over 2000), to roughly $7.1M 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. Day-to-day operation of the Institute is

handled by 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 ongoing Institute operations.

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

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: Networking (internet research), Algorithms, AI, and Speech. Consistent with this, the bulk of this report is organized along these lines, with one sub-report for each of the four groups.

Across these groups, two strong themes can be seen: Internet research and Human Centered Computational Intelligence (HCCI), a term intended to encompass a range of topics related to the use of computing for improved human-machine interfaces and human-human cooperation. Other topics of study are pursued opportunistically. Some of these efforts will ultimately grow into a larger activity, given sufficient Investigator and Sponsor interest. In particular, this year we identified two areas for likely future expansion: computational biology (starting with current work on bioinformatics in the Algorithms group, and on cognitive modeling in the AI group); and studies of IT and Society (based on a collaboration between Prof. Feldman, Prof. Castells of the Sociology Dept., and Pekka Himanen of Finland).

Two new activities that do not fit neatly into these categories are a quantitative study of games and a new collaboration with the Haas Business School at UC Berkeley.

SENIOR RESEARCH STAFF

The previous section briefly describes ICSI research in terms of major research themes and working groups. Future work could be extended to new 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 (April 2002) senior research staff members at ICSI, along with a brief description of their current interests and the Group with which the researcher is most closely associated. This is perhaps the best snapshot of research directions for potential visitors or collaborators. Not included here are a range of postdoctoral Fellows, faculty associates, 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: word sense disambiguation, computer-assisted translation, information extraction, etc.

SALLY FLOYD (NETWORKING): congestion control, transport protocols, queue management, and network simulation.

ATANU GHOSH (NETWORKING): extensible open source routing, active networks, protocols, multimedia, and operating systems.

RAMESH GOVINDAN (NETWORKING): wireless sensor networks, multicast interoperability testing, topologies.

STEVEN GREENBERG (SPEECH): Spoken language processing by humans and machines, analysis of spontaneous speech at the phonetic and prosodic levels, automatic labeling and segmentation of phonetic and prosodic material in spontaneous speech corpora, analysis of automatic speech recognition systems, computational models of auditory processing, hearing-aid design for improving speech intelligibility, auditory-visual speech processing.

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.

SRINII NARANAYAN (AI): Probabilistic models of language interpretation, graphical models of stochastic grammars, semantics of linguistic aspect, on-line metaphor interpretation, embodied rationality. Also, the role of sub-cortical structures in attentional control.

VERN PAXSON (NETWORKING): intrusion detection; Internet measurement; measurement infrastructure; packet dynamics; self-similarity.

BARBARA PESKIN (SPEECH): speech processing, including recognition and tracking of speech and speakers.

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

SCOTT SHENKER (NETWORKING): congestion control, internet topology, game theory and mechanism design, scalable content distribution architectures, and quality of service.

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

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

CHUCK WOOTERS (SPEECH): systems issues for speech processing, particularly for automatic speech recognition; universals in pronunciation modeling.

RESEARCH GROUP REPORTS

In 2001 the four core groups continued to demonstrate international leadership in their respective areas. It was a difficult year for the world of technology as we entered what appears to have been a global slowdown. While we have begun to feel some of the effects of the slowdown, ICSI has continued to expand in all aspects despite adverse conditions: industrial funding continued throughout the year; as usual, the Institute won new competitive Federal grants; the XORP project grew; and we proceeded with the groundwork for new areas in computational biology and in "Information Technology and Society", a collaborative effort between technologists and social scientists. In addition, the Finnish visitor program gradually ramped up, as did a new partnership with the Haas Business School that is includes both German and Finnish visitors.

RESEARCH GROUP HIGHLIGHTS

The following are a few key achievements for the year 2001, both in terms of 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 past year.

NETWORKING

Proposal for the Datagram Control Protocol (DCP), a new transport protocol for congestion-controlled, unreliable transfer.

A new tool, TBIT (TCP Behavior Inference Tool), for identifying the TCP behavior of web servers.

Proposal for Adaptive RED, a modified version of the RED active queue management mechanism, along with an evaluation of active queue management mechanisms.

Study of the "Constancy" of Internet path properties.

New positions: Sally Floyd elected to Internet Architecture Board, and Vern Paxson appointed Chair of the Internet Research Task Force.

Co-organized the first ACM Sigcomm Workshop on Internet Measurement.

ALGORITHMS

Intensified activity in computational molecular biology, with projects in:

- multiple sequence alignment
- peptide sequencing,
- discovery of protein sequence motifs
- sequence assembly,
- sequencing by hybridization,
- analysis of gene expression data.

Three accepted papers from the Group for RECOMB 2000, one of the two leading conferences in computational genomics.

Successful proposal process leading to a medium ITR grant from NSF on Applications of Theoretical Computer Science to the Natural and Social Sciences (with C. Papadimitriou, A. Sinclair and U. Vazirani)

AI

Significant advances in the use of Neural Grammar in courses and in applications, both linguistic and computational.

Collaboration established with the Helen Wills Neuroscience Institute, UC, Berkeley. Behavioral and imaging studies designed to explore key predications arising from the Neural Theory of Language and of SMRITI, a biologically realistic model of episodic memory formation.

New database structure and software tools completed for annotation and frame editing in the FrameNet project; also significant speedups achieved.

Visits from and collaboration with FrameNet colleagues at AT&T labs, U Penn, U of Brighton, and Vassar College. Planned cooperation with FrameNet-like projects for Spanish (Barcelona), German (U Texas), and Finnish (Tampere).

Contract received from the Army Research Institute (via CTI) to model critical thinking in the battlefield using SHRUTI, a neurally plausible model of reasoning and decision making.

New collaboration in applied language understanding with SIMS at UCB and with Stanford University, funded by ARDA.

Addition of Srini Narayanan to AI research staff (as of March 2002)

SPEECH

Best WER reduction on noisy digits at Eurospeech (with Columbia and OGI)

English-language SmartKom system (developed in collaboration with AI group and with DFKI in Germany) demonstrated at Eurospeech

Groundbreaking corpus of multichannel speech from meetings recorded, transcription nearly complete.

Speech analysis progress, including study of relation between pronunciation variation and stress, and automatic emotion (frustration) detection via prosody.

Addition of Barbara Peskin and Chuck Wooters to speech staff

New funding from DARPA and NSF for research on speech in meetings, and from Qualcomm for robust speech recognition for cell phones

Nelson Morgan appointed to IEEE Speech Technical Committee

NETWORKING GROUP

INTERNET ARCHITECTURE

HOST-BASED CONGESTION CONTROL ALGORITHMS

Because TCP is the most prevalent form of congestion control on the Internet, there is widespread consensus that any non-TCP congestion control algorithm must be *TCP-friendly* or *TCP-compatible*; that is, it must coexist with TCP without consuming an undue share of the bandwidth. ICSI researchers, along with others, have been working on an approach to TCP-friendly congestion control called *equation-based* congestion control. Equation-based congestion control (EBCC), rather than imitating TCP's window adjustment algorithm, seeks to match TCP's bandwidth usage equation in the longer run (the equation describing TCP's bandwidth usage resulting from a given packet-drop rate), while avoiding TCP's halving of the sending rate in response to a single packet drop.

This past year we have considered the implications of TCP-friendly congestion control, particularly those variants that, like EBCC, have slower responses to congestion than that of TCP. We have explored these algorithms in terms of fairness, responsiveness, and the potential for extended periods of congestion; this work [5] appeared in Sigcomm 2001.

Almost all Internet congestion control work is done in the context of TCP. However, the use of UDP is increasing rapidly, and UDP does not employ any congestion control at all. Thus, we have embarked on the design on a new protocol, the Datagram Control Protocol (DCP), in an attempt to bring congestion control to unreliable flows of unicast datagrams. DCP was first widely discussed in the IETF at a session in December 2001. However, much of the progress on DCP has occurred in 2002 and is not covered in this report.

Whilst we believe that our work on equation-based congestion control for unicast applications is now mature enough to consider standardization and real-world deployment, our research on equation-based congestion control for multicast applications is still ongoing. In addition to the problems associated with stable congestion management that emerge for unicast, with multicast there is a tension between providing rapid feedback, and preventing an implosion of feedback messages at the sender. Thus the control loop for multicast must necessarily function with larger delays. We have investigated a number of possible solutions for feedback suppression in the context of this control loop, and at the present time this work is looking rather promising. There are however still some potentially undesirable side-effects when the degree of statistical multiplexing is low and the number of receivers is large, and so this work is ongoing. Unlike with unicast congestion control, this multicast work is currently entirely simulation-based, as it is extremely hard to investigate such scalability issues in real networks. This work [39] appeared in Sigcomm 2001.

Several other developments in the field of congestion control are described in [11].

ROUTER-BASED CONGESTION CONTROL ALGORITHMS

Random Early Detection (RED) is the most widely accepted and studied form of Active Queue Management (AQM). However, the original RED design had several drawbacks, including the need to set many parameters, and the inability to determine the average queue length directly. While these failings have prompted the design of many radically different approaches to AQM, ICSI researchers sought to address these problems within the framework of the original RED design. The resulting proposal, Adative RED, appears to function well over a wide range of conditions [13], thereby suggesting that the original RED framework remains a viable approach to AQM.

While TCP-friendly congestion control is widely deployed, there is nothing in the current Internet infrastructure or architecture that enforces adherence to TCP-friendliness. ICSI researchers have made a series of proposals of how routers might protect the network (and other flows) from misbehaving flows, particularly flows that are not responsive at all. In a series of papers, a wide spectrum of approaches have been investigated that involve differing amounts of complexity and state, and that result in different levels of fairness between flows; see [25, 23, 19, 35]

However, there are many severe congestion scenarios that don't involve the misbehavior of individual flows. In particular, distributed denial-of-service attacks and flash crowds are two scenarios where links can be overwhelmed by the presence of many well-behaved (but perhaps not well-intentioned) flows. An ICSI research project on pushback and aggregate-based congestion control focuses on identifying when a rise in the packet drop rate is due to an increase in traffic from a traffic aggregate, as a subset of the traffic of the congested link. If the router is able to identify a traffic aggregate largely responsible for the traffic increase, whether the aggregate is due to malicious behavior (a DDoS attack) or merely the legitimate convergence of interest on a particular site (a flash-crowd), the router might want to preferentially drop packets from that aggregate, to protect the rest of the traffic on that link from an overall increase in then packet drop rate. Coupled with this preferential dropping at the congested router, the router might invoke pushback to request the immediate upstream routers to also drop packets from this aggregate, and this dropping might propagate recursively. This use of pushback prevents an unnecessary waste of bandwidth by packets that will only be dropped downstream. In the case of a denial-of-service attack, pushback could help focus the preferential packet-dropping on the malicious traffic within the identified aggregate [22].

IPNL

This project proposes a solution to the IP address depletion problem. Unlike IPv6, IPNL does not require any changes to the core routers; instead, it leverages the DNS (Domain Name System) to provide a routing and addressing architecture that retains the look and feel of IP to the extent possible. The key is to use FQDNs (Fully Qualified Domain Names) in an end-to-end routing mechanism that, in addition, provides benefits such as address-space separation between providers and subscribers (leading to better address-space aggregation at the provider, and better multihoming support for the subscriber), and safe mobility. Briefly, we use variable-length FQDNs as globally unique and persistent addresses to bootstrap communication between end points, and two sets of aligned and fixed-length numeric addresses to facilitate fast routing from that point onwards. One set of addresses is locally

unique and transient, while the other set of addresses comprises the usual IP address space. IPNL-routers (that exist only at the edges of the IP-network) interface between end-points which are assigned addresses drawn from the locally-unique, fast and transient address space (the first set) on one side, and the usual IP realm on the other side. Unlike other solutions, such as TRIAD, this scheme is architected to be lightweight (state at IPNL-routers only proportional to number of active local hosts, and not number of active connections) and scalable (transparent load-balancing between multiple routers exists), robust to router crashes and link failures (mechanisms for fail-over, fault-isolation, and recovery exist), and secure against a variety of attacks (such as address-spoofing, replay, and Denial of Service) against mobile hosts. This work appeared in Sigcomm 2001 [14].

MEASUREMENTS

TCP BEHAVIOR INFERENCE TOOL

There are a range of TCP congestion control behaviors in deployed TCP implementations, include Tahoe, Reno, NewReno, and Sack TCP, which date from 1988, 1990, 1996, and 1996, respectively. We recently asked the question "What fraction of the non-SACK TCP flows in the Internet use NewReno instead of Reno congestion control mechanisms? ". One reason for asking the question "Who uses NewReno TCP?" is to better understand the migration of new congestion control mechanisms to the public Internet. A second reason to ask the question is to discourage network researchers from extensive investigations of the negative impacts of Reno TCP's poor performance when multiple packets are dropped from a window of data, if in fact Reno TCP is already being replaced by NewReno TCP in the Internet. A third reason to investigate the TCP congestion control mechanisms actually deployed in the public Internet is that it is always useful to occasionally step back from our models, analysis, and simulations, and look at the Internet itself.

Inferring the congestion control behavior of a remote host can be done to some extent by actively initiating a transfer of data from a remote host over a TCP connection, and then passively monitoring the connection's congestion control responses to packet drops on congested links within the network, if in fact packets are dropped. Passive monitoring was used by Paxson in previous work. However, it is difficult to determine TCP congestion control behaviors through passive monitoring, because one has to wait for the desired pattern of packet drops to occur.

We have developed a tool, called the TCP Behavior Inference Tool (TBIT) for answering this and related questions about TCP behavior. The tool uses some of the code developed by the Sting project. TBIT allows the user to control the receipt and sending of TCP packets at the local host, thus introducing specific packet drops at the host itself. Apart from testing for the congestion control ``flavor'', TBIT is capable of testing for correctness of SACK implementation, use of timestamps, ECN capabilities and several other TCP behaviors. Both the TBIT source code and the experimental results are available from the TBIT Web Page. TBIT has already helped detect and correct bugs in Microsoft, Cisco and IBM products. This work appeared in Sigcomm 2001 [24].

CONSTANCY OF INTERNET PATH PROPERTIES

Many Internet protocols and operational procedures use measurements to guide future actions. This is an effective strategy if the quantities being measured exhibit a degree of *constancy*: that is, in some fundamental sense, they are not changing. In this project we explored three different notions of constancy: mathematical, operational, and predictive. Using a large measurement dataset gathered from the NIMI infrastructure, we then applied these three Internet path properties: loss, delay, and through-put. Our aim was to provide guidance as to when assumptions various forms of constancy are sound, versus when might prove misleading [42].

TOPOLOGY

The seminal paper by Faloutsos et al. described the power-law degree distribution of the Internet topology graph, at both the AS and router level. Ever since then, researchers have been proposing models that would explain this degree distribution. In a series of papers, ICSI researchers have analyzed and critiqued some of the most widely held beliefs about the formation of these power-law graphs. In particular, in [4] the linear preferential attachment model is compared to real data and found wanting; in [40] the general approach of assuming "criticality" as the explanation for power-laws is critiqued. An alternative explanation, which posits that the power-law AS degree distribution follows from the distribution of AS sizes is described in [37].

ICSI researchers have also investigated the impact of policy routing on path-lengths [38], and how one can best infer AS topology from various network measurements [3].

NETWORK INTRUSION DETECTION

BRO

The Bro project (LBNL and ICSI) is a network intrusion detection system. It sniffs packets coming across a network link such as DMZ or a sensitive LAN and uses an event engine to analyze the traffic and extract from it events at different levels (e.g., connection attempted; user authenticated; FTP file retrieve request; new line of Telnet output). It then determines whether the traffic is consistent with the site's policy by running the series of events as input to a script that expresses the policy in a domain-specific language. The script can maintain and modify global state, record information to stable storage, synthesize new events, generate real-time alerts, and invoke shell scripts as a form of "reactive firewall".

Bro is currently operational at several sites (ICSI, LBNL, UCB, NERSC, ESNET, JGI, Saarland University). The code is freely available.

REFLECTORS

A potentially devastating counter that distributed denial-of-service attackers can use to attempt to thwart proposed techniques for tracking back spoofed traffic flows to their source is to bounce their attack traffic off of "reflectors", i.e., any Internet host that when sent a packet will return one in reply. In this work we discuss the general problem and analyze the different types of reflectors available in the Internet to assess the degree to which we can defend against reflector attacks [26].

NORMALIZERS

A fundamental problem for network intrusion detection systems is the ability of a skilled attacker to evade detection by exploiting ambiguities in the traffic stream as seen by the monitor. We discuss the viability of addressing this problem by introducing a new network forwarding element called a traffic normalizer. The normalizer sits directly in the path of traffic into a site and patches up the packet stream to eliminate potential ambiguities before the traffic is seen by the monitor, removing evasion opportunities. We examine a number of tradeoffs in designing a normalizer, emphasizing the important question of the degree to which normalizations undermine end-to-end protocol semantics. We discuss the key practical issues of cold start and attacks on the normalizer, and develop a methodology for systematically examining the ambiguities present in a protocol based on walking the protocol s header. We then present norm, a publicly available user-level implementation of a normalizer that can normalize a TCP traffic stream at 100,000 pkts/sec in memory-to-memory copies, suggesting that a kernel implementation using PC hardware could keep pace with a bidirectional 100 Mbps link with sufficient headroom to weather a high-speed flooding attack of small packets [15].

PEER-TO-PEER CONTENT DISTRIBUTION SYSTEMS

The past few years have seen the dramatic (and surprising to many of us) rise of peer-to-peer content distribution systems. Napster and Gnutella are the most well-known and most popular examples of this genre, but new peer-to-peer content distribution systems seem to arise daily. However, most of these systems, including Napster and Gnutella, suffer from severe scalability problems.

There has been much recent research interest in how one might make such systems scalable. Several groups, including one at ICSI, have concluded that a key component of scalable peer-to-peer distribution systems is a *Distributed Hash Table* (DHT) which maps "keys" onto "values" and so any host can store a file based on its (well-known) name, and then any other host can retrieve that file just by knowing its name. Note how this differs from the web, where to locate a web site you must now only know the name of the file, but the location (the web site).

While our initial motivation is scalable peer-to-peer distribution systems, we conjecture that many other large-scale, distributed systems could utilize a DHT as a core system building block. Large scale storage management systems like OceanStore and Publius, and distributed, location independent name resolution services are but two examples.

Our work on DHTs started with the design of a *Content-Addressable Network* or CAN, which is a particular implementation of a DHT [29]}. We have followed that with studying the several other projects:

- DHTs are usually defined as an overlay network, where nodes are placed randomly in this this logical overlay. This results in each overlay hop potentially having long latency. We studied how one could guide the construction of the overlay network so that the resulting overlay hops were short [30].

- We proposed a way to use DHTs (CAN in particular) as a way to achieve scalable application-level multicast distributions [31]. Such an approach would obviate the need for specialized multicast overlays and would instead leverage the use of the more general purpose DHT.

- We have investigated DHT routing algorithms. We have developed an algorithm that has superior characteristics to previous proposals.

- We have explored the notion of a Grass-Roots content distribution network, which envisions bringing the advantages of a RAID-like approach to the web [28].

EXTENSIBLE OPEN ROUTER PLATFORM

IP router software currently differs from software written for desktop computers in that it tends to be written for a specific manufacturer's router family rather than being portable across systems from many different manufacturers. Whilst this makes sense in the case of code that interacts closely with high-speed forwarding engines, this observation also holds true for higher-level software such as routing protocols and management software. In addition, the APIs that would enable such higher-level software to be written by third parties are typically also not made public.

We are developing an open and extensible software platform for routers that might change this router software development model.

There are two main parts to such a platform:

- Higher-level routing code comprising the routing protocols, routing information bases, and management software necessary to exist in today's Internet.

- Low-level kernel code comprising an efficient forwarding path, together with the appropriate APIs to talk to routing code and to talk to additional higher-level functionality that is likely to be added later.

At the higher-level, the project needs to develop an architecture for interconnecting higher-level protocol modules in a manner that is efficient, but also modular and flexible. The APIs between these modules need to be carefully designed and well specified to allow third-parties to contribute new modules that extend the functionality of the router. It is important that such extensions can be in binary-only form so that a wide range of third-party business models are possible. It is also important that a router can use well-verified trusted routing modules at the same time as experimental third-party modules without unduly compromising the stability of the router. In this way such a router platform would stimulate early experimentation and deployment of new network functionality. For good operational reasons, this is extremely difficult in the current Internet.

At the low level the architecture needs to be capable of spanning a large range of hardware forwarding platforms, ranging from commodity PC hardware at the low end, through mid-range PC-based platforms enhanced with intelligent network interfaces, to high-end hardware-based forwarding engines. Initially we plan to focus on a PC-based hardware

platform as this allows for easy testing and early deployment, but great care will be taken to design a modular forwarding path architecture which can support some or all of the forwarding functions happening in hardware.

The primary motivation for ICSI's involvement with this is that it is becoming increasingly hard for researchers to influence the development of protocols that require router support unless we can convince a router vendor that they have a business case for the protocol in question. However we also believe that if the project is completely successful, it may result in a change in the way that commercial router software is developed, allowing more competition and resulting in lower costs and better software for ISPs.

This project was funded in 2001 by Intel and AT&T. We have recently received a large award from NSF to augment the funding for this project.

SENSORNETS

Directed diffusion is a data-centric routing mechanism for data dissemination in wireless sensor networks. In directed diffusion, low-level communication mechanisms are defined in terms of named data.
This *low-level naming*, because it eliminates the many levels of name bindings found in traditional computer networks, enables energy-efficient application designs. Low-level naming also allows intermediate nodes to *aggregate* data. This *in-network aggregation* reduces communication cost and promotes energy-efficiency. In [16], we demonstrated the performance advantages of these architectural principles through actual implementation.

In other work on diffusion routing we examined the efficacy of careful tree construction algorithms on network lifetime. Greedy trees--which promote increased sharing of paths at the expense of increased latency--allow greater data aggregation, compared to shortest-path trees, at higher network densities [18]. Finally, we proposed and evaluated a technique for monitoring large-scale sensor nets [43]. Such a monitoring capability can inform incremental deployment of sensor nodes in depleted regions of the sensor field.

MOBILITY AND QUALITY-OF-SERVICE

In 2001 we concluded the networking activity grouped under the heading "Network Services and Applications". This activity was largely conducted by visiting researchers from Spain and Germany. The activities followed the research described in our 2000 Activity report, focused primarily on mobility and quality of service**.**

USAIA: Ubiquitous Service Access Internet Architecture: Continuing the work of 2000, we developed a Petri Net to examine the behavior of mobile nodes in the case of passive resource reservations in the neighborhood of the current location. Based on this Petri Net, simulations with different parameters were investigated to understand the overall system behavior with respect to the number of admitted flows in the base stations. The goal was to maximize the probability of seamless QoS for mobile nodes in the network. Details of the examined architecture and the results can be found in [34]. The whole activity was transferred to Siemens as an internal project in co-operation with the Technical University of Berlin and it is currently part of a project funded by the German government.

Routing for mobile access scenarios: This activity examined the approach to support a hybrid operational mode for WLAN adapters. Details with respect to power control are presented in [33].

Pervasive computing: This work examined necessary components for a pervasive computing framework [2] in co-operation with University of Dresden, Germany. In addition to the development of a TCP network monitoring mechanism, a study was completed to classify terminal devices, communication systems, and user related context information for pervasive applications. Furthermore investigations were performed to get a better understanding of context representation languages. The result is an outline of a comprehensive structured context profile representation language.

IP Traffic Engineering with MPLS: This work was slightly modified from the research of the year 2000. The focus was switched on inter-area QoS routing. This investigation dealt with the question, how different domains, capable of performing intra QoS routing, can improve the network performance by perfoming QoS routing between these domains. Results are the definition of a route metric and a methodology for a constraint-based, shortest path first, routing algorithm. This algorithm was then implemented in the ns (network simulator) tool and basic simulations were performed.

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 Networking 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 was recently appointed the Chair of the IRTF. Sally Floyd was appointed to the Internet Architecture Board (IAB), which is a technical advisory board to the Internet Society and the IETF.

REFERENCES

PAPERS

[1] S. Akella, S. Seshan, S. Shenker, and I. Stoica. Exploring congestion control. preprint.

[2] S. Buchholz, S. Gobel, A. Schill, and T. Ziegert. 2001. Dissemination of mutable sets of web objects. 13th IASTED International Conference on Parallel and Distributed Computing and Systems.

[3] H. Chang, R Govindan, S. Jamin, S. Shenker, and W. Willinger. On inferring AS-level connectivity from BGP routing tables. preprint.

[4] Q. Chen, H. Chang, R Govindan, S. Jamin, S. Shenker, W. Willinger. 2002. The origin of power laws in internet topologies revisited. *Proc. IEEE Infocom 2002*. Forthcoming.

[5] D. Bansal, H. Balakrishnan, S. Floyd, and S. Shenker. 2001. Dynamic behavior of slowly-responsive congestion control algorithms. *SIGCOMM 2001*.

[6] J. Feigenbaum, C. Papadimitriou, R. Sami, and S. Shenker. Incentive-compatible interdomain routing. preprint.

[7] J. Feigenbaum, A. Krishnamurthy, R. Sami, and S. Shenker. Hardness results for multicast cost sharing. preprint.

[8] J. Feigenbaum, A. Krishnamurthy, R. Sami, and S. Shenker. 2001. Approximation and collusion in multicast cost sharing. preprint. (Abstract appears in *Proceedings of the 3rd Conference on Electronic Commerce*, 2001, pp. 253-255).

[9] J. Feigenbaum, C. Papadimitriou, and S. Shenker. 2001. Sharing the cost of multicast transmissions. *Journal of Computer and System Sciences* 63: 21-41. (Special issue on Internet Algorithms. Earlier version in *STOC '00*).

[10] S. Floyd and V. Paxson. 2001. Difficulties in simulating the Internet. *IEEE/ACM Transactions on Networking*, 9:4, pp. 392-403.

[11] S. Floyd. 2001. A report on some recent developments in TCP congestion control. *IEEE Communications Magazine*. April.

[12] S. Floyd. 2001. Simulation is crucial. Sidebar, *IEEE Spectrum*. January.

[13] S. Floyd, R. Gummadi, and S. Shenker. "Adaptive RED: An algorithm for increasing the robustness of RED's active queue management". preprint.

[14] P. Francis, R. Gummadi. 2001. IPNL: A NAT-extended Internet architecture. Sigcomm 2001.

[15] M. Handley, C. Kreibich, and V. Paxson. 2001. Network intrusion detection: Evasion, traffic normalization, and end-to-end protocol semantics. *Proc. USENIX Security Symposium 2001*.

[16] J. Heidemann, F. Silva, C. Intanagonwiwat, R. Govindan, D. Estrin, and D. Ganesan. 2001. Building efficient wireless sensor networks with low-level naming. *Proceedings of the Symposium on Operating Systems Principles*. Banff, Canada. October 2001.

[17] M. Harren, J. Hellerstein, R. Huebsch, B. Loo, S. Shenker, and I. Stoica. 2002. Complex queries in DHT-based peer-to-peer networks. In *Proc. of 1st International Workshop on Peer-to-Peer Systems* (IPTPS '02). Forthcoming.

[18] C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann. Impact of network density on data aggregation in wireless sensor networks. International Conference on Distributed Computing Systems (ICDCS-22).

[19] R.M. Karp, S. Shenker, and C.H. Papadimitriou. A simple algorithm for finding frequent elements in streams and bags. *Transactions on Database Systems*. Submitted.

[20] Q. Lv, P. Cao, E. Cohen, E. Felten, and S. Shenker. Search and replication in unstructured peer-to-peer networks. preprint.

[21] Q. Lv, S. Ratnasamy, and S. Shenker. 2002. Can heterogeneity make Gnutella scalable? In *Proc. of 1st International Workshop on Peer-to-Peer Systems* (IPTPS '02). Forthcoming.

[22] R. Manajan, S. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, and S. Shenker. Controlling high bandwidth aggregates in the network. Submitted to *Computer Communication Review*.

[23] R. Mahajan, S. Floyd, and D. Wetherall. 2001. Controlling high-bandwidth flows at the congested router. ICNP 2001.

[24] J. Padhye and S. Floyd. 2001. Identifying the TCP behavior of web servers SIGCOMM 2001.

[25] R. Pan, L. Breslau, B. Prabhakar, and S. Shenker. Approximate fairness through differential dropping. preprint.

[26] V. Paxson. 2001. An analysis of using reflectors for distributed denial-of-service attacks. *Computer Communication Review* 31(3).

[27] P. Radoslavov, R. Govindan, and D. Estrin. 2001. Topology-informed Internet replica placement. In *Proceedings of Sixth International Workshop on Web Caching and Content Distribution*. June 2001.

[28] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. Grass-roots content distribution: RAID meets the Web. preprint.

[29] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. 2001. A scalable content-addressable network. Sigcomm 2001.

[30] S. Ratnasamy, M. Handley, R.Karp, and S. Shenker. Topologically-aware overlay construction and server selection. In *Proceedings of Infocom 2002*. Forthcoming.

[31] S. Ratnasamy, M. Handley, R.Karp, and S. Shenker. 2001. Application-level multicast using content-addressable networks. In *Proceedings of NGC 2001*.

[32] S. Ratnasamy, S. Shenker, and I. Stoica. 2002. Routing algorithms for DHTs: Some open questions. In *Proc. of 1st International Workshop on Peer-to-Peer Systems* (IPTPS '02). Forthcoming.

[33] M. Sanchez. 2001. Adaptive power control for ad-hoc networks. *5th International Conference on Systemics Cybernetics and Informatics*, SCI 2001.

[34] J. Sokol and D. Tutsch. 2001. Petri net based performance evaluation of USAIA's bandwidth partitioning for the wireless cell level. PNPN Petri nets and Performance Models. September 2001.

[35] I. Stoica, H. Zhang, and S. Shenker. Self-verifying CSFQ. 2002. In *Proc. of IEEE INFOCOM'02*. Forthcoming.

[36] I. Stoica, D. Adkins, S. Ratnasamy, S. Shenker, S. Surana, and S. Zhuang. 2002. Internet Indirection Infrastructure. In *Proc. of 1st International Workshop on Peer-to-Peer Systems* (IPTPS '02). Forthcoming.

[37] H. Tangmunarunkit, J. Doyle, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. 2001. Does AS size determine degree in AS topology? *ACM Computer Communication Review*. October 2001.

[38] H. Tangmunarunkit, R. Govindan, and S. Shenker. 2001. Internet path inflation due to policy routing. In *Proceeding of SPIE ITCom 2001*. Denver, CO. 19-24. August 2001.

[39] J. Widmer and M. Handley. 2001. Extending equation-based congestion control to multicast applications. SIGCOMM 2001

[40] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. 2002. Scaling phenomena in the Internet: Critically examining criticality. In *Proceedings of Natl. Acad. Sci. USA*. Vol. 99, Suppl. 1, 2573-2580.

[41] W. Willinger, V. Paxson, R. H. Riedi, and M. S. Taqqu. 2002. Long-range dependence and data network traffic. In *Long-range Dependence: Theory and Applications.* P. Doukhan, G. Oppenheim and M. S. Taqqu, eds.. Birkhauser. Forthcoming.

[42] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker. 2001. On the constancy of Internet path properties. In *Proc. ACM SIGCOMM Internet Measurement Workshop*. November 2001.

[43] J. Zhao, R. Govindan, and D. Estrin. 2002.Residual energy scans for monitoring wireless sensor networks. 2002. IEEE Wireless Communications and Networking Conference (WCNC'02).

INTERNET DRAFTS

[44] S. Floyd, S. M. Bellovin, J. Ioannidis, K. Kompella, R. Manajan, and V. Paxson. 2001. Pushback messages for controlling aggregates in the network. draft-floyd-pushback-messages-00.txt. internet-draft, work in progress. July 2001.

[45] M. Handley, J. Padhye, S. Floyd, and J. Widmer. 2001. TCP friendly rate control (TFRC): Protocol Specification. Internet draft draft-ietf-tsvwg-tfrc-03.txt, work in progress, May.

[46] E. Kohler and S. Floyd. 2001. Profile for DCP congestion control ID 0: Single-window congestion control. draft-kohler-dcp-ccid0-01.txt, November.

[47] M. Handley, I. Kouvelas, T. Speakman, and L. Vicisano. 2001. Bi-directional protocol independent multicast (BIDIR-PIM). draft-ietf-pim-bidir-03.txt, June.

[48] S. Floyd. 2001. Inappropriate TCP resets considered harmful. draft-floyd-tcp-reset-01.txt, work in progress, June.

[49] S. Floyd and L. Daigle, Eds. 2001. IAB architectural and policy considerations for OPES. draft-iab-opes-01.txt, work in progress, October.

RFCS

[50] K. Ramakrishnan, S. Floyd, and D. Black. 2001. The addition of explicit congestion notification (ECN) to IP. RFC 3168. PROPOSED STANDARD, September.

[51] B. Whetten, L. Vicisano, R. Kermode, M. Handley, S. Floyd, and M. Luby. 2001. Reliable multicast transport building blocks for one-to-many bulk-data transfer. RFC 3048. INFORMATIONAL, January.

[52] M. Allman, H. Balakrishnan, and S. Floyd. 2001. Enhancing TCP's loss recovery using limited transmit. RFC 3042. PROPOSED STANDARD, January.

ALGORITHMS

During 2001 the Algorithms Group conducted research on Networking, Auction and Mechanism Design, Combinatorial Optimization, Computational Biology, and Probabilistic Combinatorics. The members of the Group were Richard Karp (leader), postdoctoral fellows Ernst Althaus, Eran Halperin, Robert Krauthgamer, Amir Ronen and Matthias Westermann, and several students and short-term visitors. Research was conducted jointly with faculty at UC Berkeley and at several universities in Germany and Israel. There was an especially active collaboration with the ICSI Center for Internet Research (the part of the Networking Group known as ACIRI during 2001), and some of our networking research is described in the Networking section of this report.

COMBINATORIAL OPTIMIZATION

SCIL - SYMBOLIC CONSTRAINTS IN INTEGER LINEAR PROGRAMMING

Ernst Althaus

Many combinatorial optimization problems are naturally formulated through constraints. Consider the Traveling Salesman Problem (TSP). It asks for the minimum cost Hamiltonian cycle in an undirected graph $G = (V,E)$ with edge costs. We want to select a subset S of the edges of G such that "S is a Hamiltonian cycle" and the sum of the costs of the edges in S is minimized.

Our vision is that the paragraph above (written in some suitable language) should suffice to obtain an efficient algorithm for the TSP. Efficiency is meant in a double sense. We want short development time (= efficient use of human resources) and runtime efficiency (= efficient use of computer resources). SCIL is our first step towards realizing this vision.

We propose a programming system that features *symbolic constraints* in a branch-and-cut framework. For us, an optimization problem is specified by a set of variables (ranging over real numbers), a linear objective function, and a set of symbolic constraints. A symbolic constraint is simply a subset of the set of all assignments of values to variables. An assignment is feasible if is satisfies all symbolic constraints.

SCIL has already been used for curve reconstruction [7], surface reconstruction from planar contours [3], docking with flexible side chains [5], and multiple sequence alignment.

"SOMEWHAT SATISFYING" ASSIGNMENTS

Eran Halperin

We introduce a new notion of *somewhat satisfying assignments*. In a CSP problem (Constraint Satisfaction Problem) we are given a set of variables, and a set of clauses. Each clause contains a set of variables or their negations (literals), and the goal is to find a {0,1} assignment to the variables, such that a certain boolean function of the literals in each clause will be satisfied. Given constraint satisfaction problems Π_1 and Π_2, we say that Π_1 somewhat satisfies Π_2 if there is a polynomial-time algorithm which finds a satisfying assignment for Π_1 on every instance for which Π_2 is satisfiable. For example, C.

McDiarmid has shown that there is a polynomial-time algorithm which, given a 3-colorable graph, finds a coloring with two colors, such that there is no monochromatic triangle. We characterize the set of 3-CSP problems (CSP problems with 3 literals per clause). We show that, for every pair Π_1, Π_2 of 3-CSP problems, the problem of finding a satisfying assignment for $\Pi_1$when Π_2 is satisfiable is either NP-hard or solvable in polynomial time. We further show that for a given satisfiable instance *I* of SAT, we can find a satisfying assignment to *I*, if for every clause of *I*, the number of true literals in the satisfying assignment is given. Our algorithms use semi-definite programming. We also introduce random walk based algorithms using the method of Mcdiarmid, for a subset of the problems mentioned above.

DISCRETE METRIC SPACES AND THEIR ALGORITHMIC APPLICATIONS

Robert Krauthgamer

Metric embeddings: beyond one-dimensional distortion

Metric spaces and their embeddings have recently played a prominent role in the development of new algorithms. So far, most of the emphasis has been on embeddings that preserve pairwise distances. A very intriguing concept introduced by Feige [12], makes it possible to quantify the extent to which higher-dimensional structures are preserved by a given embedding. In [18] we investigate this concept for several basic graph families such as paths, trees, cubes and expanders.

Property testing of data dimensionality.

Data dimensionality is a crucial issue in a variety of settings, where it is desirable to determine whether a data set given in a high-dimensional space adheres to a low-dimensional structure. We study this problem in the framework of property testing [14]: Given query access to a data set *S*, we wish to determine whether *S* is low-dimensional, or whether it must be modified significantly in order to have the property. Allowing a constant probability of error, we aim at algorithms whose complexity does not depend on the size of *S*.

In [19] we present algorithms for testing the low-dimensionality of a set of vectors and for testing whether a matrix is of low rank. We then address low-dimensionality in metric spaces. For l_1 metrics, we show that low-dimensionality is not testable. For l_2 metrics, we show that a data set can be tested for having a low-dimensional structure, but that the property of *approximately* having such a structure is not testable.

COMPUTATIONAL BIOLOGY

COMBINATORIAL OPTIMIZATION IN BIOINFORMATICS

The introduction of combinatorial optimization methods to the field of computational biology has proven to be a useful tool for numerous applications such as RNA secondary structure alignment, computing the fit of three-dimensional structures, and a general trace formulation of multiple sequence alignments.

MULTIPLE SEQUENCE ALIGNMENT

Ernst Althaus

Aligning DNA or protein sequences is certainly one of the dominant tools in computational molecular biology. The spectrum of methods ranges from extremely fast hashing-based methods to moderately expensive pairwise comparisons based on dynamic programming, to costly, exact multiple alignment formulations, which are either based on the natural extension of the dynamic programming paradigm, or on the application of combinatorial optimization techniques.

We have extended the formulation of the *gapped trace problem* proposed by Reinert [25] such that we can formulate a great variety of multiple sequence alignment problems, among them the weighted sum of pairs problem with arbitrary gap costs. To our knowledge this is the first algorithm that can deal with truly affine gap costs. Indeed our algorithm is independent of the choice of the gap cost function and can handle any function including convex gap costs which were proposed in several publications.

DE NOVO PEPTIDE SEQUENCING VIA TANDEM MASS SPECTROMETRY

Ernst Althaus, Richard Karp, and Ashwin A. Seshia.

Peptide sequencing via tandem mass spectrometry is one of the most powerful tools in proteomics for identifying proteins. Previous algorithms are based on a rather simple computational model. We developed a more elaborate model and give two algorithms to solve it. One algorithm is based on integer programming, the other on branch-and-bound.

FINDING PROTEIN MOTIFS VIA METRIC EMBEDDINGS

Eran Halperin, Richard Karp, and Robert Krauthgamer

Protein motifs are preserved domains in a protein sequence. Proteins with a common motif often have the same function or fold. The problem of finding protein motifs in a set of sequences containing it is well studied. The algorithms known for DNA motifs are much more satisfactory than those known for protein motifs, as the similarity between two protein sequences is more difficult to quantify. Recently, Buhler [11] suggested an approach that is based on transforming the similarity measure between amino acids (e.g. PAM or BLOSUM) into a Hamming distance between binary strings.

In this work we suggest two improvements to the problem. We first propose a metric to represent the similarity function. We justify the relation between the metric and the similarity in terms of a probabilistic model. We then compute better transformations mapping amino acid sequences to binary strings such that the Hamming distances between the strings approximate the distance between the amino acid sequences. Based on the observation that the desired transformation is essentially a finite metric that embeds isometrically into l_1 (or into a hypercube), we use techniques that are common in the area of metric embeddings (e.g. semidefinite programming) to devise algorithms for computing nearly optimal transformations (i.e. embeddings). We analyze our algorithms theoretically and evaluate them empirically.

HANDLING LONG TARGETS AND ERRORS IN SEQUENCING BY HYBRIDIZATION.

Eran Halperin

This work was initiated before coming to ICSI, and the main theoretical parts were done in Tel-Aviv. Some of the work of implementing the simulations of the algorithms, and of writing the paper, were done at ICSI.

Sequencing by hybridization is a DNA sequencing technique, in which the sequence is reconstructed using its k-mer content. This content, which is called the spectrum of the sequence, is obtained by hybridization to a universal DNA array. Standard universal arrays contain all k-mers for some fixed k, typically 8 to 10. Currently, in spite of its promise and elegance, sequencing by hybridization is not competitive with standard gel-based sequencing methods. This is due to two main reasons: lack of tools to handle realistic levels of hybridization errors, and an inherent limitation on the length of uniquely reconstructable sequence by standard universal arrays.

In this work, we deal with both problems. We introduce a simple polynomial-time reconstruction algorithm which can be applied to spectra from standard arrays and has provable performance in the presence of both false negative and false positive errors. We also propose a novel design of chips containing universal bases, that differs from the one proposed by Preparata et al. [24]. We give a simple algorithm that uses spectra from such chips to reconstruct with high probability random sequences of length lower only by a squared log factor compared to the information theoretic bound. Our algorithm is very robust to errors, and has a provable performance even if there are both false negatives and false positive errors. Simulations indicate that its sensitivity to errors is also very small in practice.

LOCAL STRUCTURE IN GENE EXPRESSION DATA

Richard Karp

The paper [8] concerns the discovery of patterns in gene expression matrices, 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 genes by comparing their expression levels in all experiments, or clustering experiments by comparing their expression levels for all genes. This work presents methods that go beyond such global approaches by looking for local patterns that manifest themselves when we focus simultaneously on a subset G of the genes and a subset T of the experiments. Specifically, we look for *order-preserving submatrices* (OPSMs), in which the expression levels of all genes induce the same linear ordering of the experiments. Such a pattern might arise, for example, if the experiments in T represent distinct stages in the progress of a disease or in a cellular process, and the expression levels of all genes in G vary across the stages in the same way.

We define a probabilistic model in which an OPSM is hidden within an otherwise random matrix. Guided by this model we develop an efficient algorithm for finding the hidden OPSM. The algorithm works very well on matrices generated according to the model. Computations on a breast cancer data set also seem to reveal highly significant local patterns.

Our algorithm can be used to discover more than one OPSM within the same data set, even when these OPSMs overlap. It can also be adapted to handle relaxations and extensions of the OPSM condition. For example, we may allow the different rows of GxT to induce similar but not identical orderings of the columns, or we may allow the set T to include more than one representative of each stage of a biological process.

THE RESTRICTION SCAFFOLD PROBLEM

Richard Karp

Most shotgun sequencing projects undergo a long and costly phase of finishing, in which a partial assembly forms several contigs whose order, orientation and relative distance is unknown. In [9] we propose a new technique that supplements the shotgun assembly data by cheap and simple complete restriction digests of the target. By computationally combining information from the contig sequences and the fragments sizes measured for several different enzymes, we seek to form a "scaffold" on which the contigs will be placed in their correct orientation, order and distance. We give a heuristic search algorithm for solving the problem and report on promising preliminary simulation results. The key to the success of the search scheme is the very rapid solution of its two time-critical subproblems that are solved precisely in linear time. Our simulations indicate that with noise levels of some 3% relative error in measuring fragment sizes, most datasets of 20 contigs can be correctly ordered, and the remaining datasets have most of their pairs of neighboring contigs correct. Hence, the technique has a potential to provide real help to finishing. Even when the target clone remains unfinished, the ability to order and orient the contigs correctly makes the partial assembly both more accessible and more useful for biologists.

PROBABILISTIC COMBINATORICS

Richard Karp

In [1] we consider a coalescing particle model where particles move in discrete time. At each time period, each remaining ball is independently put in one of n bins according to a probability distribution and all balls put into the same bin merge into a single ball. Starting with k balls, we are interested in the properties of the expected time until all balls merge into one. We derive both upper and lower bounds on this expectation and show that the expectation is a Schur concave function. The problem is related to earlier work on random walks in graphs and clustering and dispersion rates for interacting particle systems.

DATA MANAGEMENT IN COMPUTER NETWORKS

Matthias Westermann

In [21], which is a full version of [20], we present static data management strategies for computer systems connected by networks. A basic functionality in these systems is the interactive use of shared data objects that can be accessed from each computer in the system. Examples for these objects are files in distributed file systems, cache lines in virtual shared memory systems, or pages in the WWW. In the static scenario we are given read and write request frequencies for each computer-object pair. The goal is to calculate a placement of the objects to the memory modules, possibly with redundancy, such that a given cost function is minimized.

With the widespread use of commercial networks, as, e.g., the Internet, it is more and more important to consider commercial factors within data management strategies. The goal in previous work was to utilize the available resources, especially the bandwidth, as well as possible. We will present data management strategies for a model in which commercial cost instead of the communication cost are minimized, i.e., we are given a metric communication cost function and a storage cost function.

We introduce new deterministic algorithms for the static data management problem on trees and arbitrary networks. Our algorithms aim to minimize the total cost. To our knowledge this is the first analytic treatment of this problem that is NP-hard on arbitrary networks. Our main result is a combinatorial algorithm that calculates a constant factor approximation for arbitrary networks in polynomial time. Further, we present an algorithm for trees that calculates an optimal placement of all objects in X on a tree $T=(V,E)$ in time $O(|X|\cdot|V|\cdot\text{diameter}(T)\cdot\log(\text{degree}(T)))$.

AUCTION AND MECHANISM DESIGN

Amir Ronen

Amir Ronen's main research interest is the intersection between computer science and game theory or economics. This year, Amir focused on two main topics. Optimal auction design, and fault tolerant mechanism design.

OPTIMAL AUCTIONS

The design of optimal auctions is a fundamental and beautiful problem in economics and electronic commerce. In this problem, there is one object for sale, and n potential buyers. The goal is to design an optimal selling mechanism. (The dual problem of designing an optimal buying mechanism is similar.) Specifically, each buyer i, has a *privately* known value for the object. Given a distribution on the vector of buyer's values, the goal is to construct a mechanism that maximizes the seller's expected revenue (optimal auction). This mechanism has to satisfy several standard game theoretic properties, otherwise strategic manipulation by the buyers will prevent it from achieving its goal. A comprehensive introduction to this problem, along with a collection of major results can be found at [17].

For the case where the agents are independent, the optimal mechanism is well characterized [22]. Despite the importance of the problem in micro-economics, little progress has been made for the general case. Computer science has a great chance of contributing to the understanding of this problem, mainly due to its notions of approximation.

Last year, Amir showed that simple generic mechanisms, approximate the optimal mechanisms [26] In particular, he showed a 2-approximation for the general problem. This gives rise to many intriguing questions. This year, Amir focuses on proving that the optimal mechanism requires exponential time. Such a result will show that a good characterization of the optimal mechanism, which the economists are looking for, is impossible. There are several technical complications but a sketch of the proof is already prepared. (Joint work with C. Papadimitriou and F. Wu from Berkeley CS and A. Saberi who is visiting ICSI.)

FAULT TOLERANT MECHANISM DESIGN

Mechanism design is the study of protocols that aim to function well in a game theoretic environments (i.e. environments in which the participants are self interested). However, the theory does not take into account the possibility that the agents might fail to fulfill their commitments.

Together with Ryan Porter, Yoav Shoham, Moshe Tennenholts from Stanford CS, Amir made some initial steps in merging the basic concept of fault tolerance with mechanism design solutions. A paper [23] describing these results has been submitted to AAAI.

REFERENCES

[1] I. Adler, H-S. Ahn, R.M. Karp and S.M. Ross. Coalescing times for IID random variables. Random Structures and Algorithms (submitted).

[2] E. Althaus. 2001 Curve reconstruction and the traveling salesman problem. Ph.D. Thesis, Universitat des Saarlandes.

[3] E. Althaus and C. Fink. 2002. A polyhedral approach to surface reconstruction from planar contours. In *Proceedings of the Ninth Conference on Integer Programming and Combinatorial Optimization*. Lecture Notes in Computer Science. Forthcoming.

[4] E. Althaus, O.Kohlbacher, H.P.Lenhof, and P. Muller. 2000. A branch & cut algorithm for the optimal solution of the side chain placement problem. Tech report Vol.2000-1-001**,** Max-Planck-Institut fur Informatik : Forschungsbericht.

[5] E. Althaus, O.Kohlbacher, H.P.Lenhof, and P. Muller. 2000. A combinatorial approach to protein docking with flexible side-chains. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology* (RECOMB-00) Tokyo, Japan. eds. R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman**,** pp.15–24, ACM Press. N.B. Full Version submitted to *Journal of Computational Biology*.

[6] E. Althaus and K. Mehlhorn, 2000. TSP-based curve reconstruction in polynomial time. In *Proceedings of the 11th Symp. Discrete Algorithms*. pp.686-695. ACM and SIAM.
url: http://www.mpi-sb.mpg.de/ mehlhorn/ftp/TSP-curve.ps

[7] E. Althaus and K. Mehlhorn. 2001.TSP-based curve reconstruction in polynomial time. *SIAM Journal of Computing*. 31(1).

[8] E. Althaus, K. Mehlhorn, S. Naher, and S. Schirra. 2000. Experiments on curve reconstruction. In *Proceedings of the 2nd Workshop Algorithm Engineering and Experiments* (ALENEX00). Lecture Notes in Computer Science, Springer-Verlag.
url: http://www.mpi-sb.mpg.de /mehlhorn/ftp/exp-curve.ps

[8] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. 2002. Discovering local structure in gene expression data; the order-preserving submatrix problem. *Recomb 2002* Forthcoming.

[9] A. Ben-Dor, R.M. Karp, B. Schwikowski, and R. Shamir. 2002. The restriction scaffold problem. *Recomb 2002*. Forthcoming.

[10] A. Bockmayr and T. Kasper. 1998. Branch and infer: A unifying framework for integer and nite domain constraint programming, *INFORMS Journal on Computing*.10(3): 287-300.

[11] J. Buhler. 2002. Provably sensitive indexing strategies for biosequence similarity search. In *Proceedings of the 6th Annual International Conference on Computational Biology*. ACM, April 2002. Forthcoming.

[12] U. Feige. 2000. Approximating the bandwidth via volume respecting embeddings. *Journal of Comput. System Sci.*. 60(3): 510–539.

[13] J. Giesen. 1999. Curve reconstruction, the TSP, and Menger's theorem on length. In *Proc. 15th Annual ACM Symp. Comput. Geom.* 207–216.

[14] O. Goldreich, S. Goldwasser, and D. Ron. 1998. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750.

[15] E. Halperin, R. Karp, and R. Krauthgamer. 2002. Finding protein motifs via metric embeddings. Work in progress.

[16] T. Kasper. 1998. A unifying logical framework for integer linear programming and finite domain constraint programming. Ph.D. thesis. Fachbereich Informatik, Universitat des Saarlandes.

[17] P. Klemperer. 1999. Auction theory: a guide to the literature. *Journal of Economic Surveys*. pp 227–286.

[18] R. Krauthgamer, N. Linial, and A. Magen. 2001. Metric embeddings beyond one-dimensional distortion. Manuscript. November 2001.

[19] R. Krauthgamer and O. Sasson. 2001. Property testing of data dimensionality. Manuscript, July 2001.

[20] C. Krick, H. Racke, and M. Westermann. 2001. Approximation algorithms for data management in networks. In *Proceedings of the 13th ACM Symposium on Parallel Algorithms and Architectures* (SPAA) 2001: 237–246.

[21] C. Krick, H. Racke, and M. Westermann. 2001. Approximation algorithms for data management in networks. *Theory of Computing Systems*.(invited).

[22] R. B. Myerson. 1981.Optimal auction design. *Mathematics of OperationalRresearch*. 6:58–73.

[23] R. Porter, A. Ronen, Y. Shoham, and M. Tennenholts. 2002. Fault tolerant mechanism design. Submitted.

[24] F. P. Preparata, A. M. Friese, and E. Upfal. 1999. On the power of universal bases in sequencing by hybridization. *RECOMB 1999*.

[25] K. Reinert. 1999. A polyhedral approach to sequence alignment problems. Ph.D. thesis, Universitat des Saarlandes.

[26] A Ronen. 2001. On approximating optimal auctions. In *TheThird ACM Conference on Electronic Commerce* (EC01): 11–17. ACM.

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

Jerry Feldman

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. Most people today cannot program their VCRs or use more than half the functionality of their answering machines. 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(ILI) 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. 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.

The core 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, is completing 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. In 2001 we decided to orient this work towards the DAML+OIL framework, a proposed standard for the WWW.

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 and Mark Paskin). 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 2001 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. 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 Abby Wright.

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.

FRAMENET PROJECT

Charles Fillmore

The NSF-sponsored FrameNet project began in 1997 with NSF IRI-9618838 "Tools for Lexicon Building" with the goal of creating an online lexicon for English, based on **frame semantics** and supported by corpus evidence. The project is now in its second major 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.

Major software used by FrameNet includes corpus-management tools developed at IMS-Stuttgart, frame editing and sentence annotation tools developed in-house, and several types of report generation tools developed in-house, and a web interface to a MySQL-based search tool developed by Prof. Hiroaki Sato of Senshu University, Japan.

MILESTONES

Software development:
A major activity of 2001 was building the new relational database and the custom-built software for annotation and frame editing; these were substantially completed and working in time for the summer. Software development was headed by Beau Cronin, with assistance from Wendy Wooters, Carol Hays, Olya Gurevich, Josef Ruppenhofer, and Collin Baker. After the summer, bug fixes continued and new features were added, notably the ability to annotate multiple target words within a sentence, to annotate multiple, overlapping frame elements on the same piece of text, and to annotate the governors of targets even before a frame for them has been defined.

Intensive Annotation:
As envisioned in the original project budget, May-August, 2001 was a period of rapid annotation, with many students working full-time during all or part of the summer. Thanks to the new software and the summer work schedule, we were able to annotate more than 1,200 lexical units in the first year of FN2 (Aug 2000-Aug 2001), compared with 1,850 in the three years of FN1.

Release of FrameNet I Data:
The descriptions of frames and frame elements and the annotated sentences for the 1,800 lexical units from the first three years of the FrameNet project were publicly released on May 14, 2001 on the FN website (http://framenet.icsi.berkeley.edu/~framenet). This included public release of the first version of the book *The FrameNet Project: Tools for Lexicon Building*, (Johnson et al. 2000).

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.

In addition to explaining the project to many visitors at ICSI throughout the year, Prof. Fillmore publicized the work of FrameNet through a variety of invited talks: at Microsoft (Redmond, WA January 23), the ICAME conference (Louvain-la-Neuve, Belgium, May 18), University of Konstanz, Germany (July 22), at a workshop and tutorial in Iasi, Romania (August 3), and at AT\&T Labs (New Jersey, August 31).

At the North American Association for Computational Linguistics Special session on WordNet and Other Lexical Resources (Pittsburgh June 3-4), Collin Baker presented a joint paper with Charles Fillmore on annotating running text within the FrameNet paradigm and gave, together with Beau Cronin, a demonstration of the FN software tools. Staff member Dan Gildea completed his doctorate and moved to U. Penn., where he has a post-doc

working with Martha Palmer and others in her group; they are planning several types of cooperation between their projects (such as VerbNet) and FrameNet.

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.

NEAR-TERM PLANS

AUTOMATIC FRAME ELEMENT RECOGNITION

The manual analysis and annotation activities of the project are extremely labor-intensive, but it is hoped that the annotation process can be automated in two ways. For frames that have already been manually analyzed and where a substantial number of annotated sentences exist, staff member Dan Gildea has developed an algorithm that trains on manual annotations and learns from how to mark frame elements in previously unseen text [12]. This system is under further development at U Colorado (Boulder).

For frames in which annotation is not available, we are creating a system whereby staff members can define a priori rules to recognize frame elements. We hope that the manual work can then be limited mainly to approving (or if necessary editing) automatically supplied frame element labels. This system should be in place for the next round of intensive annotation in the summer of 2002.

FRAMENETS FOR OTHER LANGUAGES

Carlos Subirats, of U Barcelona (Spain), visited FrameNet from July to November 2001, and will be visiting again in 2002 and 2003. He is heading a major effort to create a multilingual version of FrameNet; In addition to work on Spanish in Barcelona, academic researchers have offered to work on German (in Erfurt, Saarbrüken, and U Texas, Austin), Italian (in Pisa) and Finnish (in Tampere), and several corporate partners have expressed interest. The Berkeley team has agreed to collaborate in this "EuroFrameNet" project on a consulting basis. Funding is being sought through the European Commission; in the meantime, Prof. Subirats is beginning work on the Spanish FrameNet project with his own funding sources; a pilot study was completed in 2001.

TESTING WITH NLP APPLICATIONS

The second phase of the project includes a commitment to test whether the performance of NLP applications can be improved by the addition of data from FrameNet. Three types of applications are being modified to make use of this data by the three FrameNet Co-PIs: information extraction (Srini Narayanan, SRI International and ICSI), question answering (Dan Jurafsky, U Colorado Boulder) and machine translation (Jean Mark Gawron, San Diego State University).

CONNECTIONIST MODELING

Lokendra Shastri

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, progress was made in developing connectionist control structures for selecting among competing actions/plans, designing a schema for planning using episodic memory, and implementing a Java-based simulator for SHRUTI.

The results of the above work are described in [11] and [23]. Carter Wendelken is a graduate student, and Maximillian Garagnani was a post-doctoral fellow visiting from The Open University, England working with Shastri.

The SHRUTI system is being applied to model critical thinking in the battlefield under a project funded by the Army Research Institute.

Ongoing research focuses on (i) augmenting SHRUTI to enable it to identify complex plans, and (ii) developing neurally plausible algorithms for learning SHRUTI structures such as types (based on observed instances), composite relational schemas (from simpler schemas), and causal structures (from temporal ordering of observed events).

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 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, biologically grounded explanations for well-known behavioral findings about human memory were identified. These include the fan-effect and the list-strength effect. It is significant that no attempt was made to model these behavioral findings in the development of the model; the explanations of these phenomena arise directly from the biologically grounded architecture and structure of the model. Several behavioral predictions stemming from SMRITI were also identified.

Several behavioral and imaging studies have been designed to test the model's predictions in collaboration with cognitive neuroscientists and psychologists at UC, Berkeley and UC Davis.

Results of the above work are reported in [13-18] and were presented at CNS'01, the annual Computational Neuroscience Meeting, Asilomar, CA.

REFERENCES

[1] B. Bergen. Probability in phonological generalizations: Modeling optional French final consonants. In Alan Yu et al. eds., *Proceedings of the 26th Annual Meeting of the Berkeley Linguistics Society*. Berkeley Linguistics Society. Berkeley, CA.

[2] B. Bergen. Ramifications of phonology-syntax interactions for phonological models. In *Proceedings of the 24th Annual Penn Linguistics Colloquium*. Pp. 27-36.

[3] B. Bergen and N. Chang. 2000. Spatial schematicity of prepositions in Neural Grammar. Fifth Conference on Conceptual Structure, Discourse, and Language, Santa Barbara, CA.

[4] B. Bergen and N. Chang. Embodied Construction Grammar in Simulation-Based Language Understanding. In J. Ostman and M. Fried eds., *Construction Grammar(s): Cognitive and Cross-Language Dimensions*. Johns Benjamins. Forthcoming.

[5] B. Bergen and N. Chang. 2001. Semantic agreement and construal. 2001 Meeting of the Linguistic Society of America. Washington, D.C.

[6] H. C. Boas. 2001. Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja eds., *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK.

[7] N. Chang and R. Hunter. 2001. Reference resolution in child-directed speech: A deep semantic approach. International Workshop on Reference and Coherence in Discourse; Formal, Functional and Cognitive Approaches. Utrecht, The Netherlands.

[8] N. Chang and T. Maia. 2001. Grounded learning of grammatical constructions. 2001 AAAI Spring Symposium on Learning Grounded Representations.

[9] C. J. Fillmore and C. F. Baker. 2001. Frame Semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*. Held in conjunction with the NAACL Annual Meeting, CMU. Pittsburgh, PA.

[10] C. J. Fillmore, C. Wooters, and C. F. Baker. 2001. Building a large lexical databank which provides deep semantics. In B. Tsou and O. Kwong, eds. *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, Hong Kong.

[11] M. Garagnani, L. Shastri, and C. Wendelken. 2001. A connectionist model of planning via back-chaining search. In *Proceedings of PLANSIG 2001*.

[12] D. Gildea and D. Jurafsky. 2000. Automatic labeling of semantic roles. In *ACL 2000: Proceedings of ACL 2000*, Hong Kong.

[13] L. Shastri. 2002. Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences* 6(4): 162-168. In press. Available at http://www.icsi.berkeley.edu /~shastri/psfiles /ShastriTicsEM02.htm

[14] L. Shastri. 2002.. A computationally efficient abstraction of long-term potentiation. *Neurocomputing*. In press. Available at http://www.icsi.berkeley.edu/~shastri/psfiles /ShastriNCltp02.pdf

[15] L. Shastri. 2001. A computational model of episodic memory formation in the hippocampal system. *Neurocomputing*. 38-40:889-897. (Also appears in Bower, J. ed. *Computational Neuroscience: Trends in Research*. 2001, Elsevier, Amsterdam).

[16] L. Shastri. 2001. Biological grounding of recruitment learning and vicinal algorithms in long-term potentiation. In *Emergent neural computational architectures based on neuroscience*, J. Austin, S. Wermter, and D. Wilshaw, eds. Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence Springer-Verlag, Berlin. Pp. 348-367.

[17] L. Shastri. 2001. From transient patterns to persistent structure: A model of episodic memory formation via cortico-hippocampal interactions. Submitted. Available at http://www.icsi.berkeley.edu/~shastri/psfiles/shastri*pdf*.

[18] L. Shastri. 2001. Episodic memory trace formation in the hippocampal system: a model of cortico-hippocampal interactions. Technical Report 01-004. International Computer Science Institute, Berkeley, CA.

[19] L. Shastri, S. Chang, and S. Greenberg. 1999. Syllable detection and segmentation using temporal flow neural networks. *Proceedings of the XIVth International Congress of Phonetic Sciences, ICPhS 99*. 3:1721-1724. San Francisco, CA.

[20] L. Shastri, D. Grannes, S. Narayanan, and ,J. Feldman In press. A Connectionist encoding of parameterized schemas and reactive plans, in *Hybrid Information Processing in Adaptive Autonomous Vehicles*, G.K. Kraetzschmar and G. Palm (Eds.). Lecture Notes in Computer Science. Springer-Verlag, Berlin.

[21] C. Wendelken, and L. Shastri. 2000. Probabilistic inference and learning in a connectionist causal network. *Proceedings of Neural Computation 2000*. Berlin.

[22] C. Wendelken. 2002. The role of mid-dorsolateral prefrontal cortex in working memory: a connectionist model. *Neurocomputing*. In press (Presented at the Tenth Annual Computational Neuroscience Meeting. Pacific Grove CA. July 2002).

[23] C. Wendelken and L. Shastri. 2002. A structured connectionist agent architecture: I. Combining belief and utility. Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Fairfax, VA. August 2002. Forthcoming**.**

SPEECH PROCESSING

During 2001, the Realization group was re-named as the Speech group, better reflecting our current focus on speech processing technologies. This year's activities varied widely over this domain, from work on improved front-end processing of the speech signal to back-end analysis of attributes such as emotion in automatically processed speech. Along a different dimension, projects ranged from highly speculative explorations to involvement in actual system prototyping.

The year's Speech efforts were headed by senior research staff members Jane Edwards, Steven Greenberg, Hynek Hermansky (OGI and ICSI), Nelson Morgan, Barbara Peskin, Elizabeth Shriberg (SRI and ICSI), Andreas Stolcke (SRI and ICSI), and Chuck Wooters. Dan Ellis also continued his work with us while primarily being a faculty member at Columbia University. Of this group, Chuck Wooters and Barbara Peskin were 2001 additions; Chuck moved over from his work with the AI Group (FrameNet), and Barbara joined the Senior Research staff after a decade in the speech technology industry with Dragon Systems. As always, major contributions were also made by our team of students, postdoctoral Fellows, and visitors, and through collaborations with other speech groups.

The sections below describe a number of the year's major activities in the areas of Speech Signal Modeling, Speech Recognition, and Speech Analysis. In addition, we include a fourth section describing our Meeting Recorder project, which, though still in its early stages, is already providing a rich testbed for all three of our speech focus areas. While this listing includes many of our most significant projects, it is by no means exhaustive, but should provide a useful overview of the activities in which we have been engaged this year.

SPEECH SIGNAL MODELING

The group was involved in a number of projects focused on processing and modeling of speech signals in order to furnish information to be incorporated into systems for recognition of spoken language and other related tasks. A common theme of these efforts was to provide processing that was robust to noise, whether from high-noise environments such as in-car dictation or highly reverberant environments such as far-field recordings in conference rooms.

NOISE-ROBUST PROCESSING FOR CELLULAR STANDARDS

Performance of automatic speech recognition (ASR) is affected by a communication network that limits the available frequency range of speech and may introduce errors, a problem particularly severe for mobile communications. The effects of the communication network may be alleviated in distributed speech recognition (DSR), which performs some of its operations, namely parameterization of the speech signal, in a handset and the rest of its processing in the network. In this way, extracted speech parameters can be protected by error-correcting code for transmission over the network. To ensure interoperability of ASR systems provided by various manufacturers, the feature extraction needs to be standardized, and the European Telecommunications Standards Institute (ETSI) is now attempting to establish such a standard in its Aurora group, which includes participation of practically all major telecommunications companies.

ICSI, as one of the world leaders in research on acoustically robust speech processing, welcomed the opportunity to test our approaches in industrial applications through the Aurora project. The project is interesting to us because it does not allow for any modifications of the pattern matching or language modeling components of the recognizer and therefore requires that we focus entirely on extraction of noise-robust features. In 1999, we initiated a project addressing Aurora's goals, in collaboration with the OGI School of Oregon Health and Sciences University in Portland, Oregon, and partnering with Qualcomm, Inc., whose ETSI membership enabled our participation in these activities.

From the ICSI site, Dan Ellis and Hynek Hermansky participated in this research, collaborating with the group at OGI. Our first system, presented to Aurora in 2000, was based on the original idea of data-guided feature extraction (pioneered by our group since 1997), included our then-new Feature Net technique, the Modulation-filtered Spectrogram (MSG) features developed at ICSI, and the Temporal Patterns (TRAP) technique from OGI, and achieved the best performance among all participants. We thus ensured our place among the six participants (Nokia, Motorola, French Telecom, Alcatel, Ericsson, and ICSI/OGI/Qualcomm) who advanced to the next run.

In early 2001, we presented a second version of our algorithm, designed to work on multiple languages (the first systems had been tested only on American English digit sequences). The multilingual aspect of the new Aurora focus prevented use of some of the more aggressive approaches to speech feature extraction. The ICSI team of Stephane Dupont, Carmen Benitez and Hynek Hermansky, along with a group at OGI (including Sunil Sivadas and Pratibha Jain), based our submission mainly on data-guided RASTA filtering and partly on data-guided Feature Nets [3]. One novel aspect of the algorithm was the use of a neural net based voice activity detector that proved to be highly effective. This algorithm placed second in the competition, the best algorithm being the result of a joint effort of French Telecom and Alcatel. At about the same time, Qualcomm became an ICSI sponsor with the immediate goal of funding the Aurora-related research.

During 2001, the Aurora team (further supported by Barry Chen and David Gelbart, doing basic research aimed at robustness in realistic environments) continued efforts to further improve the data-guided RASTA filtering and on a more thorough study of the Feature Net and TRAP approaches. In addition, there was also a significant push towards research on explicit noise suppression techniques, which were missing from our second submission. Further, Chuck Wooters joined the team on a part-time basis to work on large vocabulary continuous recognition of noisy speech, a task that was added to the new evaluation, supplementing the original noisy digits work. The ICSI team closely collaborated with OGI on the task (Sunil Sividas from OGI was a summer intern at ICSI, Stephane Dupont visited OGI several times, and Hynek Hermansky holds positions at both ICSI and OGI). The team also worked closely with Hari Garudadri at Qualcomm, who provided significant help in steering our solution towards industrial needs and interfacing with the ETSI Aurora committee.

The current algorithmic solution splits processing between the hand-held terminal and the network. The hand-held part includes the neural net based voice activity detector, the modified Wiener filter for the suppression of additive noise, the data-guided RASTA filtering, and split vector quantization based feature coding with error correction. The

network side contains recursive mean and variance compensation, TRAP based Feature Net, dynamic feature calculation, frame extrapolation, and final elimination of non-speech analysis vectors. The algorithmic delay of the whole feature extraction module is below 220 ms. On the Aurora task (American English TI_DIGITS corrupted by various noises; SpeechDAT-Car Italian, Finish, German and Danish digits; and noise-corrupted large vocabulary continuous speech from Wall Street Journal data), it achieves about the same performance as the baseline system on the well-matched task (where the training and test data come from similar environments) and improves performance significantly on mismatched data (where the test data come from different environments from the training data).

Outside of the framework of the ETSI evaluations per se, we have also continued our work on Tandem modeling for the Aurora (noisy digits) task within the RESPITE project, as described in last year's report. Our latest system, using deltas and normalization between the first-stage posterior-estimation neural network and the second-stage distribution-modeling Gaussian mixture model, gave the greatest reduction in overall word error rate for the multi-condition training case out of more than 20 papers presented at the Eurospeech Special Event on the Aurora task, in Aalborg, September 2001 [7].

LONG-TERM LOG SPECTRAL MEAN SUBTRACTION

When speech is recorded using a far-field microphone, the effects of reverberation (along with air absorption and microphone response) create a channel response that distorts the speech spectrum and degrades speech recognition performance. If the window functions used for Fourier analysis are sufficiently long and smooth, the received signal spectrum is approximately equal to the product of the speech spectrum and the channel spectrum. Thus, assuming the channel does not vary over time, we can attempt to remove the effect of the channel, along with any constant-over-time portion of the speech spectrum, by removing the time average of the signal spectrum.

If we do this by subtraction in the logarithmic spectral domain, we are following the same principle as the common technique of cepstral mean subtraction. However, removing the effect of reverberation requires much longer-term analysis windows than are typically used for cepstral mean subtraction. In [1], the authors noted this and measured the effect of the subtraction method on speech recognition performance using long-term (two seconds long) analysis windows. They found that it improved ASR performance in reverberant environments, but they performed their tests using only simulated reverberation.

We used tabletop microphone recordings to evaluate this method in real far-field conditions [8]. This work, which used a system trained on the TI_DIGITS corpus and evaluated performance on far-field recordings of digit strings collected as part of our Meeting Recorder project (described below), confirmed that the method can significantly improve ASR performance. Following this, we began work on the use of the subtraction method in an interactive system, to be evaluated on the SmartKom information kiosk system. A key difference in the interactive situation is that there are increasing amounts of data available for calculating the time average as an interaction session progresses.

SPEECH RECOGNITION

The 2001 effort included several projects directly focused on speech recognition, both on fundamental modeling questions toward improving recognizer performance and on implementing working systems in the field. In addition to the items below, further work on speech recognition is reported in the Meeting Recorder section.

RECOGNITION USING PARTIAL INFORMATION TECHNIQUES

As part of the European Project RESPITE, we have continued development of novel techniques for recognition of speech in situations when significant portions of the signal are rendered unobservable by added masking interference. The basic principle of this approach is missing-data recognition, which provides an accurate and effective basis for statistical inference -- given an accurate mask defining which parts of the signal are 'good' and which are corrupt. In normal conditions, however, this mask is not known and must first be deduced from the signal.

One simple approach is to assume the interference is of the form of a constant background noise, estimate its level in each frequency band by looking at the gaps between words (i.e. the temporal energy minima), then labeling any energy significant above this level as target. This approach falls down badly, however, when the interference has a dynamic variation in energy -- for instance, mechanical clatter or competing voices.

The approach taken by the 'multisource decoder' [2] is to treat the assignment of signal fragments as target or interference as simply another aspect of the recognition decoder search. Thus, multiple alternative mask hypotheses are pursued in parallel, with the final output representing the word sequence at the 'present data mask' that gave the maximum-likelihood fit to the original speech models. Within this framework, the toughest problem is limiting the search space by organizing the signal into regions that can safely be assumed all to have arisen from the same source. To this end, we use a variety of techniques including computational auditory scene analysis to construct source fragments on the basis of common onset, periodicity, and continuity.

SPEECH RECOGNITION FOR INFORMATION KIOSKS AND PORTABLE DEVICES

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 [20]. 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 porting various other language dependent components from their original German version to English. Working on SmartKom at ICSI were postdocs Thilo Pfau and Robert Porzel, graduate student David Gelbart, and Andreas Stolcke. The project was also able to enlist the assistance of the EDU speech understanding project at ICSI and its partner EML in Germany.

In 2001, ICSI was principally involved in assembling the first full English-language version of the SmartKom demonstration system. With assistance from the German partner DFKI, we adapted the language understanding and language generation modules to English, mainly by writing appropriate English semantic grammars for the SmartKom domain. We also ported the speech synthesis module, based on Edinburgh's FESTIVAL software, by

integrating English voices. The English recognizer was tuned to the demonstration domain, which proved challenging in the absence of any English training data for the SmartKom domain. We solved this problem by combining translations of German-language SmartKom dialogs and web resources to train a suitable language model. The English SmartKom system was successfully demonstrated at the Eurospeech conference [21].

The next version of the system will address tourist information tasks (English-speaking visitors wanting to get around the city of Heidelberg), and will be based on native English speech data. For this task we collected data inhouse at ICSI, specially designed to elicit pronunciation variants of German names spoken by native English speakers, one of the main research problems facing the project. The data collection used the ICSI meeting recorder facility, allowing us to record with multiple microphones, which will support future research on acoustic robustness in this domain.

David Gelbart's research on spectral subtraction for robustness to reverberation was in part supported by the SmartKom project. The results of that work, recently published in [8] are currently being integrated into the SmartKom recognizer.

SPEECH ANALYSIS

A significant component of the Speech work this year had a focus on better understanding how various types of information -- such as phone identity, stress accent, emotional content -- are encoded in the speech signal and how such features can be automatically extracted and tagged. Other analyses have sought to understand what "causes" errors in existing speech recognition systems and to automatically discover categories of speech sound attributes which can aid speech modeling.

ARTICULATORY-ACOUSTIC FEATURE CLASSIFICATION

A novel framework for automatic articulatory-acoustic feature extraction has been developed for enhancing the accuracy of place- and manner-of-articulation classification in spoken language. The "elitist" approach focuses on frames for which multilayer perceptron, neural network classifiers are highly confident, and discards the rest. For the high-confidence frames, it is possible to achieve a frame-level accuracy of 93% for manner information on a corpus of American English sentences passed through a telephone network (NTIMIT). Place-of-articulation information is extracted for each manner class independently, resulting in an appreciable gain in place-feature classification relative to performance for a manner-independent system. Comparable classification performance for the elitist approach is evidenced when applied to a Dutch corpus of quasi-spontaneous telephone interactions (VIOS). The elitist framework provides a potential means of automatically annotating a corpus at the phonetic level without recourse to a word-level transcript, and could thus be of utility for developing training materials for automatic speech recognition and speech synthesis applications, as well as aid the empirical study of spoken language. [4,5,22]

AUTOMATIC LABELLING OF STRESS ACCENT

Stress accent is an integral component of many languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation. Multilayer perceptrons (MLPs) were trained on a portion of the Switchboard corpus of conversational telephone speech in order to automatically label the material with respect to stress accent (partitioned

into five levels of accent, from completely unaccented to heavily accented). The automatically derived labels are highly concordant with those of human transcribers (79% concordance within a quarter-step of accent level and 97.5% concordant within a half-step of accent level). In order to achieve such a high degree of concordance it is necessary to include features pertaining not only to the duration and amplitude of the vocalic nuclei, but also those associated with speaker gender, syllabic duration and most importantly, vocalic identity. Such results suggest that vocalic identity is intimately associated with stress accent in spontaneous American English, thereby providing a potential foundation with which to model pronunciation variation for automatic speech recognition (see below). [11]

ANALYSIS OF PRONUNCIATION VARIATION AND ITS RELATION TO STRESS ACCENT

There is a systematic relationship between stress accent and vocalic identity in spontaneous English discourse (as examined in the Switchboard corpus of telephone dialogues). Low vowels are much more likely to be fully accented than their high vocalic counterparts. And conversely, high vowels are far more likely to lack stress accent than low or mid vocalic segments. Such patterns imply that stress accent and vowel height are bound together at some level of lexical representation. Vocalic duration also appears to be an important acoustic cue associated with stress accent, and the association between vowel height and accent level is most clearly observed in this dimension, particularly for diphthongs and the low, tense monophthongs.

There is also a systematic relationship between stress accent and pronunciation variation in non-nucleic constituents of the syllable.

With respect to syllabic onsets, neither heavily nor lightly accented syllables exhibit a significant amount of pronunciation deviation in segments. However, unaccented syllables manifest a significant proportion of pronunciation deviations from canonical. The most common deviation is segmental deletion, which most commonly occurs when associated with words such as "them," "him" and "her," where the onset is deducible through context. The onset of "the" is frequently deleted for similar reasons. The other common form of onset deviation pertains to insertion of segments, of which the alveolar ([dx]) and nasal ([nx]) flaps, the glottal stop ([q]) and the glides [w] and [y] are the most common variety. Such segmental insertions are usually associated with some form of junctural demarcation, delineating a boundary separating an unaccented (or lightly accented) syllable from a more heavily accented precursor.

The principal effect of stress accent on syllable codas pertains to the frequency of segmental deletion (there is virtually no impact of accent on the frequency of substitutions or insertions). Fully two-thirds of the coda deletions are associated with just three segments - [t], [d] and [n], irrespective of accent weight. The heavier the accent the less likely a (canonical) coda segment will be deleted.

Such patterns imply that stress accent and syllabic articulation are inextricably bound together, and this knowledge could be used to improve pronunciation models for speech applications. Incorporation of stress-accent information into pronunciation models provides a potential means of significantly improving ASR performance beyond what is currently possible using lexical representations composed solely of phonetic-segment sequences. Stress accent can be used to interpret the acoustic signal in a manner that accommodates a variety

of insertion, deletion, and substitution phenomena commonly encountered in spontaneous discourse without significant expansion of the recognition lexicon. Moreover, such an approach is likely to minimize the mismatches that occur between stored lexical representations and the phonetic characterization of the signal performed during the recognition process through accommodation of the acoustic and pronunciation variation systematically governed at the level of the syllable. [10,11,13]

LINGUISTIC DISSECTION OF AUTOMATIC SPEECH RECOGNITION SYSTEM PERFORMANCE

A diagnostic evaluation of five Switchboard-based recognition systems was conducted in order to ascertain whether word-error patterns are attributable to a specific set of linguistic factors. Each recognition system's output was converted to a common format and scored relative to a reference transcript derived from phonetically hand-labeled data. This reference material was analyzed with respect to roughly forty acoustic, linguistic and speaker characteristics, which in turn, were correlated with recognition-error patterns via decision-trees and other forms of statistical analysis. The most consistent factors associated with superior recognition performance pertain to accurate classification of phonetic segments, articulatory-acoustic features, and syllable structure, as well as the complexity and sophistication of each system's pronunciation models. This analysis was compared with that of a comparable analysis performed the previous year on the recognition output of eight separate Switchboard systems (for a different set of data). In general, the analyses yielded similar results across years, although one important analysis, performed in 2000 was not possible to do in 2001, due to the absence of stress-accent label material for that year's evaluation material. This work was reported at NIST's Workshop on Large Vocabulary Speech Recognition in May 2001, and in [9].

AUTOMATIC LEARNING OF SPEECH SOUND ATTRIBUTES

Linguists have developed articulatory and phonetic features to describe and classify speech sounds. For example, Distinctive Features [14] are categorizations of speech along expert-derived qualities that characterize one phone as different from the next. Unfortunately, these features are often a caricature of what is actually present in the speech signal itself. This work begins to address the problem of finding the characteristics that differentiate speech sounds, by using automatic machine learning techniques.

Currently, this work has focused on examining patterns learned by neural networks whose parameters are optimized for discriminating between phones. These patterns come from the learned input-to-hidden unit weights of neural networks. The weights can be viewed as specifying a hyperplane in the input space that offers a degree of separation between classes, or, alternatively, as a pattern that causes a particular hidden unit to activate. The combination of activations among hidden units serves to optimally separate classes of inputs. By examining these hidden unit patterns, we can automatically discover important intermediate categories or attributes in speech that help to discriminate between phones.

This project is still in an early and highly exploratory stage. To date, we have trained a set of neural networks, multilayer perceptrons (MLPs) with a single hidden layer. The training material for these MLPs consists of long-context windows (typically 101 frames, derived using a 25 ms analysis window and 10 ms step-size) of cube-root compressed critical-band energies from speech drawn from the TIMIT corpus of phonetically-balanced, hand-labelled

sentences. For each critical band, a single MLP is trained to learn phone probabilities. This approach is modeled on that employed in Sharma and Hermansky's neural TRAPS [12].

When looking at the patterns learned by these independently trained critical-band MLPs, striking similarities emerge among the patterns learned for each of the critical bands. We have clustered the hidden unit weights and have seen that some clusters have members from all critical bands, while other clusters appear in a select set of critical bands. Further analysis will be done to explore the effect of each of these clusters on recognition, in particular the link between these patterns and their use in discriminating between phones. Additional experiments will be performed on neural nets trained with noisy and reverberant data, with the aim of seeing how the clustered hidden unit patterns change to help maintain robust performance in the presence of noise.

AUTOMATIC EMOTION DETECTION VIA PROSODY

In a joint project with SRI, we have investigated the use of prosody for the automatic detection of emotion in Communicator dialogs [17]. Prosody is a key indicator of emotion, as is known from work on synthesis, but there is less known about how to recognize emotion automatically in natural speech. The Communicator corpus is a DARPA-funded multi-site collaboration consisting of human-computer dialogs over the telephone. In each call, a person interacted with a fully automatic dialog system in the air travel domain, to try to make flight plans. The goal of our work was to detect points at which the user became annoyed or frustrated with the system. The ability to detect frustration -- in this corpus as well as in many other potential applications (e.g. customer service) -- would allow a system to adapt dialog strategies to the detected state, or even to refer the unhappy caller to a human operator. A longer-term goal of the project is to model emotion in conversational speech, such as in the ICSI Meeting corpus.

Five Berkeley students were hired to help develop the labeling system and label the data. We found this to be an inherently complex and difficult task, due to factors such as the inherent continuum between neutral and frustrated speech, the question of absolute versus speaker-normalized judgments, and the presence of many short utterances (like "no"). We also found it necessary to independently code for emotion and level of hyperarticulation. The resulting database encodes five basic emotions: "neutral", "annoyed", "frustrated", "tired/disappointed", and "amused/surprised". Each utterance was labeled independently by two or more labelers. In a second "consensus" pass, the two most experienced labelers re-labeled all cases of original disagreement with a "consensus" label.

The automatic modeling used decision trees to predict human original-agreed consensus labels on an unseen test set, using a variety of prosodic and other features. Our prosodic features included pause, duration, pitch (raw f_0, stylized with piecewise linear fits), energy, and spectral tilt. Features were normalized for the speaker in two modes: using all data and using only the first five utterances in a call (the latter to allow for an online model). We also included nonprosodic features, such as the utterance count and whether the utterance was a repeated request for the same information.

Results revealed that automatic prosody-based models agreed as well (or even better!) with human consensus labels, than independent human judgments agreed with each other. Main confusions in both cases were between "neutral" and "annoyed" or "annoyed" and

"frustrated", as would be expected since this is an inherent continuum. Machine agreement with human labels was much higher for those cases in which humans had originally agreed, and also much higher when only trying to distinguish "frustrated" utterances from any other type. Prosodic features, even when normalized using only the first five utterances in a call, were found to be useful predictors. For example, a high normalized pitch during the longest normalized vowel, high normalized durations or a low speaking rate, high energy, high utterance count, and repeated utterances were all associated with greater frustration. Interestingly, frustration was not well predicted by any of the pronunciation-style features, such as hyperarticulation or pausing between words -- this type of speech appears to vary independently of emotion, occurring often in both emotional and neutral speech.

THE MEETING RECORDER PROJECT

Beginning in 2000, ICSI initiated a project to collect recordings from conventional multiparty meetings. While the automatic transcription and analysis of meetings serves as a useful target application in its own right, the collection effort has been designed to provide data to support a diverse range of technologically challenging projects on the way to such an ambitious end goal, including recognition of highly spontaneous speech, processing for far-field microphone data, dialogue modeling, speaker tracking, and automatic summarization and search. One of the ICSI conference rooms was outfitted with custom hardware and software supporting up to 16-channel recordings, permitting the recording of each meeting participant over close-talking headset microphones, as well as collection of far-field data via a number of tabletop microphones. In addition to the meetings themselves, we are also collecting digit strings read by each meeting participant. This permits far-field microphone research on a more constrained task, in parallel with our efforts to tackle the harder recognition problem of spontaneous speech while using the more conventional headset microphones.

To date, we have collected over 80 hours of Meeting data and plan to complete our targeted 100-hour collection by mid-2002. This work has been supported by contracts from DARPA for corpus preparation and from NSF for "mapping" meeting content and interactions. Our effort is part of a larger program with partner sites University of Washington and Columbia University. The corpus should also support work in collaboration with various European initiatives, which have also targeted the Meeting task. A status report on the ICSI project was presented in [15].

While this project is still in its early stages, largely focusing on data collection and (human) transcription and on corpus construction at this time, it is already beginning to fuel a number of explorations, and we look forward to its supporting a rich collection of research efforts in future. In addition to the far-field microphone experiments reported in the Speech Signal section above, efforts to date have included:

TRANSCRIPTION AND ANNOTATION

As a first stage in the processing of the Meeting data, we are creating word-level transcriptions along with notation of significant non-lexical events, such as laughter and coughing, and of turn-level time-marking, in order to produce transcripts that can be used to train and evaluate automatic speech transcription systems. Spontaneous meeting data presents a number of challenges to the transcriber that traditional prepared-speech corpora

do not: how to handle disfluencies, overlaps and interruptions; how to punctuate utterances that may be truncated or abandoned and do not necessarily respect the usual written-language conventions; etc. Jane Edwards, whose expertise lies in the transcription of discourse [6], has designed a set of transcription conventions for the project and is supervising a team of students who are transcribing the material. They are assisted by an in-house modification of the publicly-available Transcriber tool, enhanced to support multiple audio channels in parallel, and by "rough draft" transcriptions created by an external transcription agency in order to speed the in-house processing. In addition to this word-level transcription, a subset of meetings is now being annotated with greater mark-up for features such as dialogue acts and disfluencies, in order to support efforts in discourse modeling and other higher-level structures.

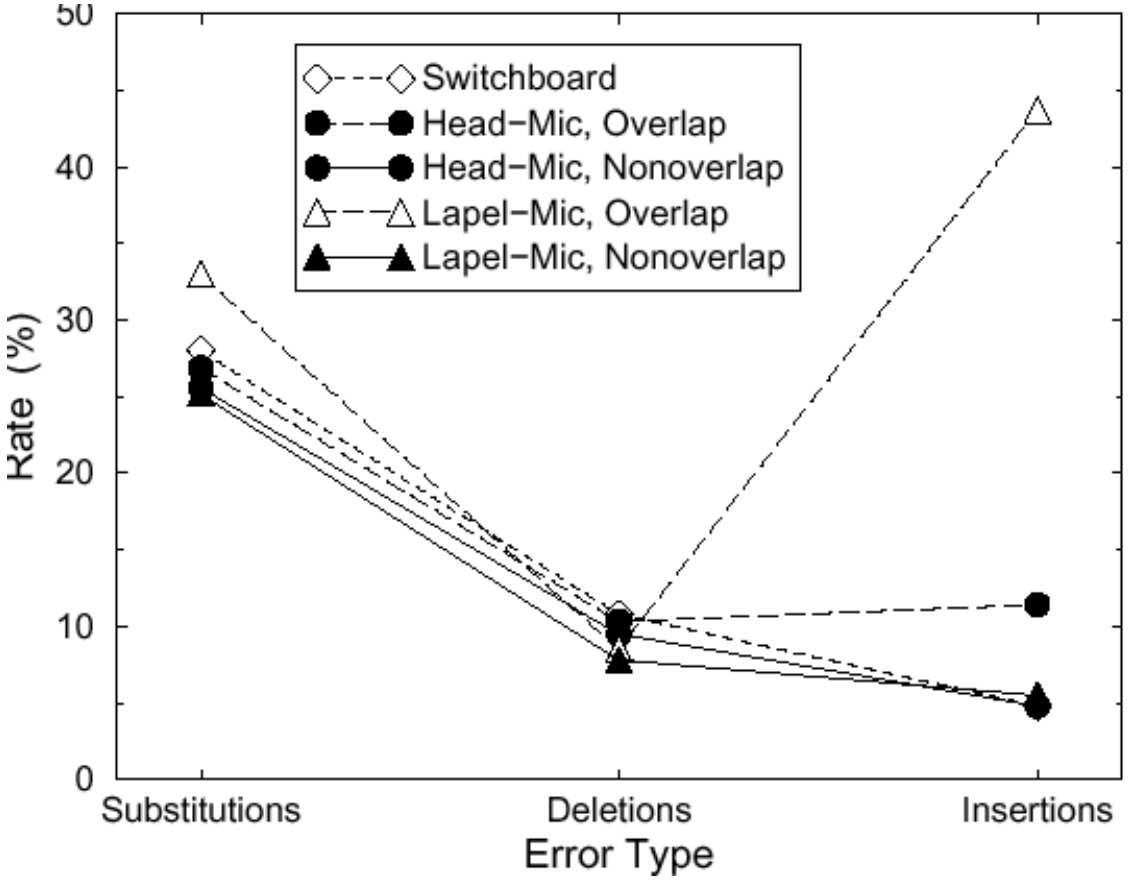
SPEECH ACTIVITY DETECTION

Because any given speaker is speaking for only a small portion of the meeting, locating regions of speech activity is an important task, both for speeding the human transcription process and for the efficiency and accuracy of automatic speech recognition. Since each speaker is recorded on his own channel via close-talking microphone, we had expected the identification of speech intervals to be straightforward. However, simple energy-based approaches yielded unacceptably high error rates in speech/nonspeech labelling, largely due to cross-talk from neighboring speakers, excessively loud breath noise from inexperienced microphone wearers, and wide variation in channel characteristics. More sophisticated approaches, such as echo cancellation via Block Least Squares, suffered from rapid changes in channel coupling due to head movements of the participants. We therefore developed our own system for multi-channel speech activity detection. Briefly, the system uses hidden Markov modeling of "speech" and "nonspeech" states via mixtures of Gaussians, and incorporates feature normalization and cross-correlation processing to increase channel independence and to detect and reduce cross-talk. Greater detail is provided in [16]. Using this system we were able to markedly reduce the frame-level speech/nonspeech classification error and bring the word error rate for automatic recognition of Meeting data to within 10% (relative) of the performance on hand-segmented data. This represents a dramatic improvement in both speed and accuracy, compared to processing the unsegmented version.

AUTOMATIC SPEECH RECOGNITION

For our first foray into automatic recognition of Meeting data, we concentrated on data from the close-talking microphones, in order to focus on speech and dialog characteristics of the data. Of the standard available corpora for training speech recognizers, the speaking style in meetings seemed most closely to resemble that of the casual telephone conversations of the Switchboard corpus, so we used a Switchboard-trained large-vocabulary recognizer from SRI, and downsampled the Meeting data to match the telephone bandwidth. We found that for data from most of our microphones, and for native speakers, performance of the SRI recognizer trained on Switchboard was about as good on Meeting data as on Switchboard data (see figure below and [18]). Although there are clear dialog-level differences (for instance, verbal backchanneling such as "uh-huh" is rarer in meetings than in telephone conversations), the overall result suggests a high degree of similarity in speaking style (language and pronunciation) between Switchboard and Meetings. This finding is important to the speech recognition community, because it suggests that government funding efforts on Switchboard data (which has had a long history, and significant data) are likely to be

largely transferable to ASR in the Meeting task. In addition to this finding, we also made the important observation that there is a significant degradation in performance when using lapel rather than head-mounted microphones, due to a much higher frequency of insertion errors. The degradation is exacerbated by the presence of background speakers during regions of speaker overlap, which (as described below) are frequent in meetings. For this reason we subsequently stopped using lapel microphones in our collection.



SPEAKER OVERLAPS

Another line of studies examined overlapping speech, i.e. temporal regions in which more than one speaker is vocalizing [18, 19]. We found a high rate of overlap in meetings: in some meeting types, 20% of words were overlapped and over 50% of speech "spurts" (continuous regions of "talk") were at least partially overlapped. The rate depended on the meeting type, with more "democratic" meetings having higher rates. A further interesting result was found when we compared rates of overlap in multiparty face-to-face meetings to those in telephone conversations: the rate of overlap was roughly the same, suggesting overlap is an inherently common characteristic of conversation, not just of meetings. Such a result has implications for the speech community, which has typically assumed that additional people cause additional overlap. Interestingly also, we found that within the class of telephone conversations (comparing Switchboard and Call Home English) speech between two strangers had no less overlap than speech between family or friends -- a finding that contradicts most current assumptions about the role of familiarity in dialog.

USE OF PROSODIC FEATURES FOR AUTOMATIC DETECTION OF SPEECH EVENTS

Finally, we conducted a large number of experiments examining the role of both lexical and prosodic features in characterizing important events in meetings, including punctuation, dialog acts, backchanneling, and overlapping speech [18, 19]. We developed a database of automatically extracted prosodic features based on extraction regions determined by forced alignment of the speech stream to the transcription. (We are currently extending this to the case of recognizer hypotheses, rather than manual transcripts, for a "fully automatic" system.) The features include duration, stylized pitch (f_0), pause, energy, and other features, normalized in various ways for the speaker. We found that such features are useful in automatic detection of punctuation in meetings -- for example in finding sentence boundaries or locations of disfluencies. They provide added value over using the words alone, even when assuming correct words. Interestingly, we also found that for the prediction of speaker overlap points, language features were not of much use, but prosodic features (both of the overlap-er and of the overlap-ee!) were useful, suggesting such features could play an important role in automatic meeting understanding.

REFERENCES

[1] C. Avendano, S. Tibrewala, and H. Hermansky. 1997. Multiresolution channel normalization for ASR in reverberant environments. *Proc. Eurospeech-97*, Rhodes, Greece. September 1997.

[2] J. Barker, M. Cooke, and D. Ellis. 2001. Combining bottom-up and top-down constraints for robust ASR: The multisource decoder. Workshop on Consistent and Reliable Acoustic Cues CRAC-2001. Aalborg, Denmark. September 2001.

[3] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivadas. 2001. Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks. *Proc. Eurospeech-01*, Aalborg, Denmark. September 2001.

[4] S. Chang, S. Greenberg, and M. Wester, 2001. An elitist approach to articulatory-acoustic feature classification. *Proc. Eurospeech-01*, Aalborg, Denmark. September 2001.

[5] S. Chang, M. Wester, and S. Greenberg. 2002. An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. Submitted to *Computer Speech and Language*. (expanded and considerably enhanced version of two Eurospeech papers).

[6]J J. Edwards. 2001. Chapter 17: The Transcription of Discourse. 2001. In *The Handbook of Discourse Analysis*, D. Shiffrin, D. Tannen, and H. Hamilton, eds. Oxford: Blackwell, pp. 321-348.

[7] D.P.W. Ellis, and M. Reyes. 2001. Investigations into Tandem acoustic modeling for the Aurora task. *Proc. Eurospeech-01*. Aalborg, Denmark. September 2001.

[8] D. Gelbart and N. Morgan. 2001. Evaluating long-term spectral subtraction for reverberant ASR. *Proc. Automatic Speech Recognition and Understanding*. Madonna di Campiglio, Italy. December 2001.

[9] S. Greenberg. From here to utility - Melding phonetic insight with speech technology. 2001. *Proc. Eurospeech-01*. Aalborg, Denmark. September 2001.

[10] S. Greenberg, H.M. Carvey, and L. Hitchcock. 2002. The relation of stress accent to pronunciation variation in spontaneous American English discourse. In *Proc. International Conference on Speech Prosody 2002*, Aix, France. April 2002.Forthcoming.

[11] S. Greenberg, S. Chang, and L. Hitchcock. 2001. The relation between stress accent and vocalic identity in spontaneous American English discourse. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*. Red Bank NJ. October 2001.

[12] H. Hermansky and S. Sharma. 1998. TRAPS - Classifiers of temporal patterns. *Proc. ICSLP '98*. Sydney, Australia. November 1998.

[13] L. Hitchcock and S. Greenberg. 2001. Vowel height is intimately associated with stress accent in spontaneous American English discourse. *Proc. Eurospeech-01*. Aalborg, Denmark. September 2001.

[14] R. Jakobson, G.M.C. Fant, and M. Halle. 1952. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. MIT Press. Cambridge, MA..

[15] N. Morgan, D. Baron, J. Edwards, D.P.W. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. 2001. The Meeting Project at ICSI. *Proc. Human Language Technology 2001*. San Diego CA.. March 2001.

[16] T. Pfau, D.P.W. Ellis, and A. Stolcke. 2001. Multispeaker speech activity detection for the ICSI Meeting Recorder. *Proc. ASRU-2001*. Madonna di Campiglio. Italy. December 2001.

[17] E. Shriberg, A. Stolcke, and J. Ang. 2001. Automatic detection of annoyance and frustration in Communicator dialogs. talk presented at the DARPA ROAR Workshop. Orlando FL. November 2001.

[18] E. Shriberg, A. Stolcke, and D. Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. Eurospeech-01*. Aalborg, Denmark. September 2001.

[19] E. Shriberg, A. Stolcke, and D. Baron. 2001. Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*. Red Bank NJ. October 2001.

[20] SmartKom: Dialog-based Human-Technology Interaction by Coordinated Analysis and Generation of Multiple Modalities. http://www.smartkom.com .

[21] W. Wahlster, N. Reithinger, and A. Blocher. 2001. SmartKom: Multimodel communication with a life-like character. *Proc. Eurospeech-01*. Aalborg, Denmark, September 2001.

[22] M. Wester, S. Greenberg, and S. Chang. 2001. A Dutch treatment of an elitist approach to articulatory-acoustic feature extraction. *Proc. Eurospeech-01*, Aalborg, Denmark. September 2001.

OTHER PROJECTS AT ICSI

As noted earlier in this report, we plan to expand our efforts in Computational Biology (currently represented by components in the AI and Algorithms Groups). We also will be starting a new program focused on Information Technology and Society. In 2001, however, two other activities already began that are significant developments, but which fall outside the scope of the principal efforts of the four Groups: the study of games (especially Go); and a new collaboration between the Haas Business School and ICSI international visitors to study the interaction between Information Technology and business. Brief descriptions of these efforts follow.

STUDIES OF QUANTITATIVE GO, AND SOME OTHER GAMES

Combinatorial game theory is concerned with two-person perfect-information games, especially those classes of positions for which winning strategies can be stated explicitly, or at least proved to exist. The powerful mathematical methods are most successful when applied to games whose positions often decompose into "sums". The many examples of such games include Nim, Dots and Boxes, Hackenbush (best played with colored chalk and erasures), Domineering (played with dominoes on a checkerboard), Konane (popular in ancient Hawaii), Amazons (invented less than fifteen years ago, but which has attracted a substantial following on the Internet), and Go (a popular Asian board game dating back several thousand years, and which supports nearly 2,000 active professionals today).

In most of these games, a mathematically defined "temperature" provides a numerical measure of the value of the next move. The extension of this notion to loopy positions, such as kos in Go, appeared in "Games of No Chance" in 1996. A subsequent extension, called "Environmental Go", includes a stack of coupons in addition to the Go board. The Berkeley game theory group, including Prof. Elwyn Berlekamp, collaborator Bill Spight, and PhD student Bill Fraser, was joined by Prof. Teigo Nakamura from Kyushu Institute of Technology in Fukuoka, Japan, who visited ICSI from April 2001 until March 2002. For one week in December 2001, we were joined at ICSI by Prof. Martin Mueller from U. Alberta and Prof. David Wolfe from Gustavus Adolphus College in St. Peter, Minnesota.

We have been working on the development of methods and techniques which allow us to get rigorous analyses of the last 50 to 100 moves of some professional games, and we not infrequently discover fatal mistakes. This has led to fruitful collaborations with professional Go players, including Jujo Jiang and Neiwei Rui.

Both are 9-dans, the highest rank in professional Go. He became famous for his star role leading China to a victory over Japan at a famous international match in the late 1980s. She has won the world's women's championship every year that it has been held, and in spring 2000, in Seoul, Korea, she became the first woman ever to win one of the major professional Go title matches in open competition. Jiang and Rui have played several "Environmental Go" demonstration games for the explicit purpose of generating more very high-level data for our research group. The most recent such game was an all-day match on January 25, 2002 at the North American Ing Foundation's Go Center in Menlo Park, CA.

Since human chess guru Gary Kasparov was defeated in a famous match by IBM's computer chess program Deep Blue, the game of Go has become the main focus of AI game-player program developers. Their progress has been very slow. As many as 50 programs now compete with each other annually in international tournaments. These efforts have been going on for several decades, inspired, at least in part, by a potential $1,000,000 prize offered by the Ing Foundation to the first program that can beat their world's "junior" (under 18) champion. No program came anywhere close, and the offer expired last year. The best Go program has a ranking in the range of 7 kyu to 5 kyu. Virtually all 50 teams writing these programs include several players rated 1 or 2 dan or better. None of the programs can yet offer serious competition to any of the programmers, even with a handicap of several stones. And Ing's junior World Champion is usually considerably stronger than any of the programmers.

We believe the main reason for this discrepancy is that the paradigm used by all good programs for playing chess or checkers fails to work for many other games, including Go. In chess and checkers, one can quickly count the material on the board, and use that as the basis of a "static evaluation function". This provides an estimate of the "score", which estimates who is ahead, and by how much. One then looks several moves ahead in the game tree and uses various search techniques to back up this function to obtain reasonably accurate assessment of the positions after various possible next moves. All current Go programs use the same technique, but, relative to chess and checkers, all current static evaluation functions for Go positions are so bad that even when minimaxed back a few moves, they often perform very poorly.

Our current work avoids this problem by focusing on endgames. The static evaluation function is correct because it is the final score. The mathematics of combinatorial game theory allow us to track these scores back through much larger numbers of moves. We can now do precise analyses of some positions which are still more than 100 moves from the end of the game, sometimes using only handwritten calculations and no computer search at all! (Typical Go games last about 250 to 300 moves from beginning to end.) Without some very major new breakthroughs (which we consider unlikely), we have no illusions of using these methods for analyzing the first 70% of any real game, but we are hopeful that our methods might someday lead to programs which would enable professional endgame commentaries of unprecedented breadth and accuracy, just as large chess endgame databases have revealed many truths not formerly known to international grandmasters.

We think the time may now be ripe for new efforts to combine modern mathematical game theory with alpha-beta pruning and other traditional AI minimax search techniques.

REFERENCES

[1] E Berlekamp. 1996. The economist's view of combinatorial games. In R. Nowakowski, editor, *Games of No Chance: Combinatorial Games at MSRI,* pages 365-405. Cambridge University Press,.

[2] E Berlekamp and D. Wolfe. 1994. *Mathematical Go: Chilling Gets the Last Point*. AK Peters, Wellesley, MA. 1994

[3] T. Nakamura and E. Berlekamp, "Analysis of Composite Corridors" http://www.icsi.berkeley.edu/ftp/pub/techreports/2002/tr-02-005.pdf

BUSINESS AND IT

In order to further the international collaboration and to examine the international IT and telecommunication environment from a business perspective, ICSI, the Fisher Center for the Strategic Use of IT (FCSUIT) at the Haas School of Business at UC Berkeley, the Center for Digital Technology and Management (CDTM) in Munich, Germany; and the Telecommunication Business Research Center (TBRC) in Finland have established a new research collaboration. The respective Centers have many similarities in their mission statements, while simultaneously providing unique complementary areas of expertise and cultural background, and all centers are integral part of leading academic institutions. Finnish participation in this program is part of the recently established *Finland Berkeley Program in Information Society and Technology*. This new partnership provides the opportunity for international research collaboration and joint outreach activities. In the fall of 2001 five research visitors from CDTM and two researchers from Finland joined this new initiative at ICSI to collaborate with the their colleagues at the Haas School of Business at UC Berkeley. Whenever appropriate ICSI, FCSUIT, CDTM, and TBRC will from time to time organize joint workshops and symposia for the promotion of the program and the dissemination of the research results.

The current focus of this international collaboration is to establish an International Technology and Strategy Forum to explore the impact of emerging technologies in an international context:

- Business opportunities created by new technologies
- Strategic deployment of new technologies
- Organizational change and management of new technologies
- Public policy and societal issues surrounding technology

Within this general context particular focus areas will be selected to investigate new trends and emerging technologies. The first such theme is "Business and Technology Trends in Wireless," an area particularly appropriate for international interdisciplinary investigation. Different competing wireless technology standards and vastly different regulatory policies in different countries have yet to lead to a uniform infrastructure for global business services. Incidentally this is also an area where European countries (especially the Scandinavian Countries) and several Asian countries are leading the world in developing and deploying wireless infrastructure along with the creation of novel business services.