ICSI

INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

International Computer Science Institute Activity Report 2004

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198 USA
phone: (510) 666 2900 (510) fax: 666 2956 info@icsi.berkeley.edu http://www.icsi.berkeley.edu

Principal 2004 Sponsors

ARDA
DARPA
Finnish National Technology Agency (TEKES)
German Academic Exchange Service (DAAD)
German Ministry of Research and Technology (BMBF)
IM2 National Centre of Competence in Research, Switzerland
Microsoft
National Science Foundation
Qualcomm
Spanish Ministry of Education and Science (MEC)

Affiliated 2004 Sponsors

BT
Cisco
DOE
DHS
European Union (via Sheffield University)
HP
Infineon
NASA

Corporate Officers

Board of Trustees

2004 Visitors in Sponsored International Programs

Name	Country	Group	Affiliation
Konsta Koppinen	Finland	Speech	TEKES
Tuomo Pirinen	Finland	Speech	TEKES
Jusso Rantala	Finland	Haas	TEKES
Pauli Ristola	Finland		TEKES
Pasi Sarolahti	Finland	Networking	TEKES
Pertti Tormala	Finland		TEKES
Rene Beier	Germany	Algorithms	DAAD
Jens Gramm	Germany	Algorithms	DAAD
Mesut Guenes	Germany	Networking	DAAD
Till Nierhoff	Germany	Algorithms	DAAD
Till Tantau	Germany	Algorithms	DAAD
Alberto Amengual	Spain	AI	MEC
Xavier Anguera	Spain	Speech	AMI
Fernando de Arriaga	Spain	AI	MEC
Javier Cardona	Spain	Networks	MEC
Marc Ferras	Spain	Speech	AMI
Javier Macias	Spain	Speech	MEC
Sira Palazuelos	Spain	AI	MEC
Carmen Pelaez	Spain	Speech	MEC
Pedro Ruiz	Spain	Networking	MEC
Carlos Subirats	Spain	AI	MEC
Francisco Valverde	Spain	AI	MEC
Frantisek Grezl	Switzerland	Speech	AMI
Vincenzo Pallotta	Switzerland	AI	IM2

DAAD: Deutscher Akademischer Austauschdienst
TEKES: Finnish National Technology Agency
MEC: Ministerio de Educacion y Ciencia
AMI: Augmented Multi-party Interaction
IM2: Interactive Multimodal Information Management, National Centre of Competence in Research, Switzerland

Contents

Part I
INSTITUTE OVERVIEW

The International Computer Science Institute (ICSI) is one of the few independent, nonprofit basic research institutes in the country, and is affiliated with the University of California campus in Berkeley, California. ICSI was started in 1986 and inaugurated in 1988 as a joint project of the Electrical Engineering and Computer Science Department (and particularly 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 research program. The goal of the Institute continues to be the creation of synergy between world-leading researchers in computer science and engineering. This goal is best achieved by creating an open, international environment for both academic and industrial researchers.

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 faculty. ICSI currently has a major focus on two areas: Internet research, including Internet architecture, related theoretical questions, open source routing, and network security; and Human Language Technology, including both speech and text processing. Additionally, there are comparatively small but significant efforts in thoretical computer science and algorithms for bioinformatics, video encoding, and studies of the Information Society.

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. The current Director of the Institute is Professor Nelson Morgan of the UC Berkeley Electrical Engineering faculty.

1 Institute Sponsorship for 2004

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

US Federal funding comes from a range of grants to support research Institute-wide. Most of this funding comes from the National Science Foundation and DARPA. International support in 2004 came from government and industrial programs in Germany, the Ministry of Science and Technology in Spain, the National Technology Association of Finland, and the Swiss National Science Foundation (through the Swiss Research Network IM2). Additional support comes from the European Union. Industrial support in

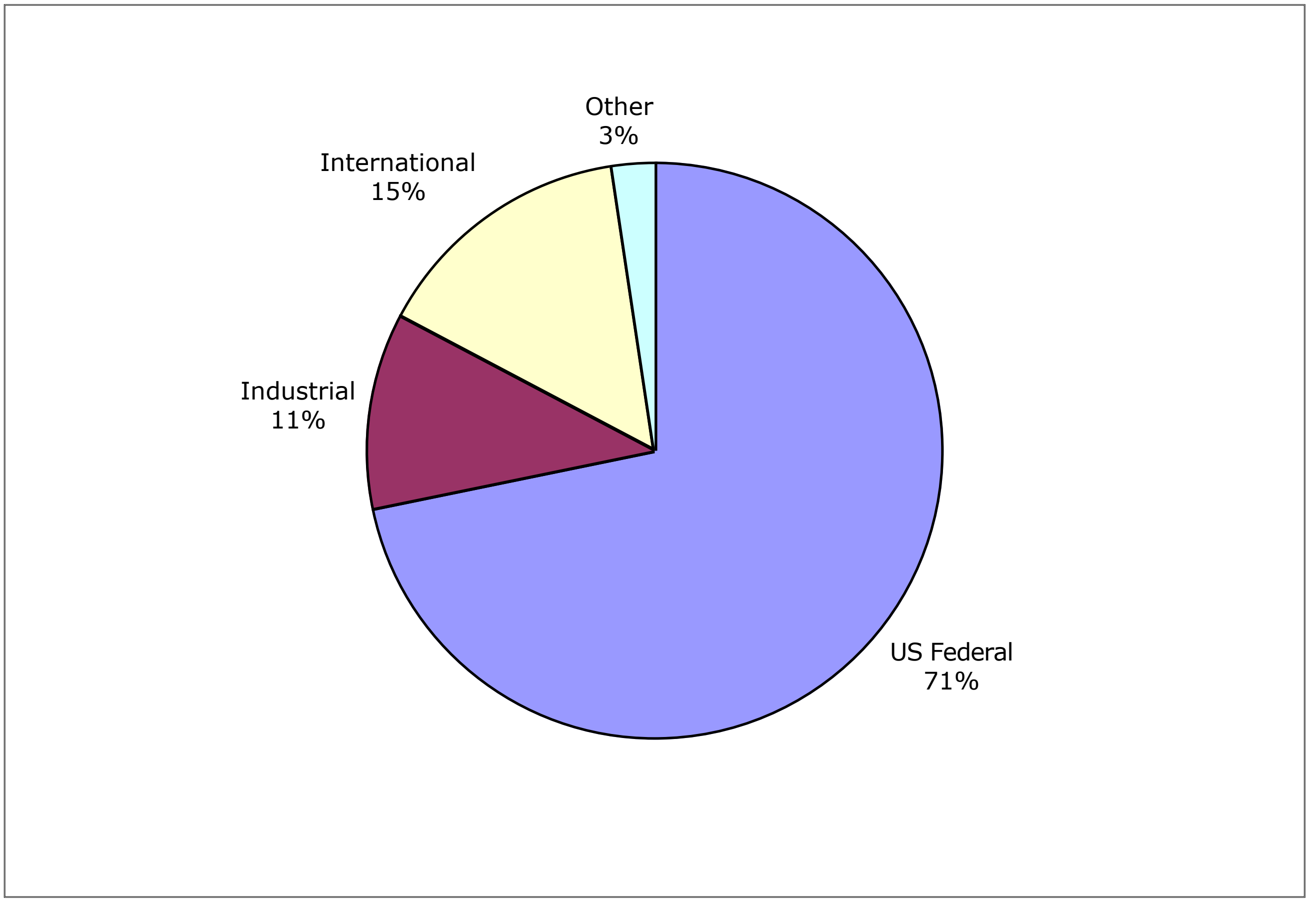


Figure 1: Distribution of sources of ICSI revenue for 2004.

2004 was primarily provided by Qualcomm and Microsoft, with additional sponsorship from British Telecom, Infineon, and Cisco.

ICSI's budget increased in 2004 (about 6% over 2003), to roughly $9.5M for the year.

2 Institutional Structure of ICSI

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

2.1 Management and Administration

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

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

We also have recently hired a new Associate Director, Dr. Marcia Bush, formerly of Xerox. She is responsible for institutional development, particularly with regards to industrial partnerships.

2.2 Research

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

Across all of these activities, there is a theme: scientific studies based on the growing ubiquity of connected computational devices. In the case of Networking studies, the focus is on the Internet; in the case of Speech and AI, it is on the interfaces to the distributed computational devices. The Algorithms group continues to develop methods that are employed in a range of computational problems, but increasingly focuses on problems in computational biology.

2.2.1 Senior Research Staff

The previous paragraphs briefly described the clustering of ICSI research into major research themes and working groups. Future work could be extended to new major areas based on strategic Institutional decisions and on the availability of funding to support the development of the necessary infrastructure. At any given time, though, ICSI research is best seen as a set of topics that are consistent with the interests of the Research Staff. In this section, we give the names of the current (March 2005) senior research staff members at ICSI, along with a brief description of their current interests and the Research Group that the researcher is most closely associated with. This is probably the best snapshot of research directions for potential visitors or collaborators. Not shown here are the range of postdoctoral Fellows, visitors, and graduate students who are also key contributors to the intellectual environment at ICSI.

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

Jerome Feldman (AI): neural plausible (connectionist) models of language, perception and learning and their applications.

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

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

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

Eran Halperin (Algorithms): Computational biology, computational aspects of population genetics, combinatorial optimization, algorithm design.

Richard Karp (Algorithms and Networking): mathematics of computer networking, computational molecular biology, computational complexity, combinatorial optimization.

Paul Kay (AI): analyzing the data from the World Color Survey, which gathered color naming data in situ from 25 speakers each of 110 unwritten languages from 45 distinct language families, in order to (1) assess whether cross-language statistical universals in color naming can be observed and (2) measure the degree to which the boundaries of color categories in individual languages can be predicted from universal focal colors.

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

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

John Moody (Algorithms): machine learning, multi-agent systems, statistical computing, time series analysis, computational finance.

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

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

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

Manny Rayner (Speech): Spoken dialogue systems, grammar-based language modeling, dialogue management, speech understanding, identification of cross-talk, evaluation methodologies. (With NASA Ames Research Center, Geneva University and Xerox Research Center Europe)

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).

Nicholas Weaver (Networking): Worms and related malcode; automatic intrusion detection and response; hardware accelerated network processing.

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

Part II
Research Group Reports

1 Research Group Highlights

The following are a selection of key achievements in our research groups for the year 2004, both in group development and in research per se. Although not a complete listing and, by necessity, quite varied given the different approaches and topics of each group, it should nonetheless give the flavor of the efforts in the ICSI community for the last year.

1.1 Networking

- XORP made its first public release in July 2004, and will soon be used for ICSI's connection to the Internet. XORP has gained extensive visibility in the community (including an article in Business Week) and is now being ported to Windows.

- The Center for Internet Epidemiology and Defenses (CIED), led by Stefan Savage of UC San Diego and Vern Paxson of ICSI, was funded by NSF. CIED aims to combat epidemic-style attacks on the Internet.

- Sally Floyd was the chosen to be the recipient of the 2005 IEEE Internet Award for "contributions to the Internet architecture, particularly in the areas of congestion control, traffic modeling, and active queue management."

1.2 Algorithms

- Eran Halperin, whose haplotyping program Hap (developed while at ICSI as a postdoc in 2002) has been widely adopted by the genetics and medical research communities, returned to ICSI as a Research Scientist.

- Roded Sharan, Richard Karp, and other colleagues developed a widely available computational method providing strong statistical evidence for many previously unobserved protein complexes, pathways, functions, and interactions.

- John Moody's group continued its developkment of Stochastic Direct Reinforcement algorithms.

1.3 Artificial Intelligence

- ICSI teamed up with Stanford and the University of Texas, Dallas to win the highly competitive ARDA sponsored AQUAINT Phase II award for building the next generation of semantically based Question Answering Systems.

- Paul Kay and Terry Regier demonstrated the presence of color naming universals in 110 unwritten languages using statistical analyses on the World Color Survey database.

- The first Global FrameNet meeting took place at ICSI in October, 2004, attended by researchers from ICSI, Saarbrücken, Barcelona, Tokyo, Austin, and Sweden. There are now active FrameNet efforts in four languages and growing worldwide interest.

- Lokendra Shastri won a DARPA award to develop the Episodic Memory module for a multi-site project to build a Perceptive Cognitive Assistant that Learns (CALO).

1.4 Speech

- The new projects awarded late in 2003 (particularly the EU project AMI, and the NSF-funded effort in speaker recognition) ramped up to full efforts. AMI also included a new visitor program, and we received our first 3 visitors.

- In collaboration with partner SRI, we fielded successful systems for metadata extraction, meeting speech recognition, and conversational telephone speech recognition for official NIST evaluations.

- The Dialog Acts annotations were released through LDC as an extension to the Meeting Corpus.

- A NASA team including ICSI researcher Manny Rayner completed the spoken dialog system Clarissa, and it was delivered to the International Space Station on Christmas Day, 2004. ICSI also began work on a new spoken language system as part of SmartWeb, a German project to provide spoken access to the Semantic Web.

1.5 Other Activities

- We completed the implementation of a working video compression system which, among other features, optimally allocated rate between distributed coded data and Forward Error Correction codes (FEC) to take advantage of FEC's error prevention capacity.

- The Berkeley Foundation for Opportunities in Information Technology" (BFOIT), centered at ICSI, partnered with a premiere mentoring program called Sage Fellows, a former UC program.
- We developed the Communities of Practice Environment (CoPE), a novel software platform for supporting cooperative effort among formal and informal groups of people who may be separated in time, space, and language.

2 Networking

2.1 Extensible Open Router Platform

XORP: Network researchers face a significant problem when deploying software in routers, either for experimentation or for pilot deployment. Router platforms are generally not open systems, in either the open-source or the open-API sense. The eXtensible Open Router Platform (XORP) attempts to address these issues. Key goals are extensibility, performance and robustness. We aim for XORP to be both a research tool and a stable deployment platform, thus easing the transition of new ideas from the lab to the real world. In the past year a release was made that is stable enough to use in production environments. To demonstrate this stability ICSI will be deploying XORP routers internally. It seems likely that XORP will be adopted commercially in the coming year.

2.2 Internet Congestion Control

DCCP: Historically, the great majority of Internet unicast traffic has used congestion-controlled TCP, with UDP making up most of the remainder. UDP has mainly been used for short, request-response transfers, like DNS and SNMP, that wish to avoid TCP's three-way handshake, retransmission, and/or stateful connections. Recent years have seen the growth of applications that therefore use UDP in a different way. These applications, including RealAudio, Internet telephony, and on-line games such as Half Life, Quake, and Starcraft, share a preference for timeliness over reliability, and use UDP instead of TCP. This growth of long-lived non-congestion-controlled traffic, relative to congestion-controlled traffic, poses a real threat to the overall health of the Internet. To address this, we are continuing the design of a new protocol, Datagram Congestion Control Protocol (DCCP), that combines unreliable datagram delivery with built-in congestion control. DCCP is currently in the final stages of standardization in the IETF.

Congestion Control in High BER Networks: One of the lingering problems faced by transport protocols in networks with high corruption-based loss rates is that congestion control is invoked on each loss under the assumption that the loss was caused by contention for resources in routers. This assumption is conservative in its effort to avoid congestion collapse. Further, the assumption is generally correct in wireline networks. However, in wireless networks where packets are corrupted on a regular basis the assumption does not hold. This work breaks down into three parts. The first is understanding the theoretical bounds on the performance implications of corruption-based loss. The second component involved a study of the impact on standard TCP of various BERs in a real satellite environment (data taken over NASA's ACTS satellite). Finally, a set of mechanisms to improve the performance in the face of high bit-error rates was introduced. The goal of this last component is to untangle the cause of losses by using an accurate total loss estimate made by the transport sender with an estimate of the corruption-based loss rate provided by the intermediate nodes in the network (routers, wireless basestations, etc.). Using these two loss rates the transport sender is able to provide a congestion response that is both appropriate to the current level of congestion while not being hindered by corruption-based losses.

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

Reducing RTOs in TCP: TCP retransmissions are triggered by either information in (duplicate) acknowledgments returned by the receiver or via a retransmission timeout (RTO). Using the RTO for loss recovery is costly from a performance standpoint. However, when the sending rate is low TCP relies on the RTO for most of its retransmissions. We have been exploring the space of possible mitigations for this problem. The two approaches we have investigated are inducing additional acknowledgments from the receiver or trigger retransmissions on a smaller number of hints in the ACK stream from the receiver. We are investigating a proposal called Smart Framing which slices data into smaller packets when there are few packets on the network and also Early Retransmit, which is designed to trigger retransmits on fewer signals from the receiver in order to circumvent the costly RTO.

Bursting in Transport Protocols: TCP's congestion control algorithms naturally send small bursts of segments into the network periodically. A re-occurring question is whether there should be some form of limit imposed on the size of these bursts. We conducted two investigations to shed light on this question. First, we conducted a theoretical evaluation of several burst mitigation strategies (both new approaches and previously proposed schemes). In addition, we conducted a study of network traces from several networks to assess the impact naturally occuring bursts have on TCP connections. We find that small scale bursting is present in most traffic, but that small scale bursting has very little performance impact. On the other hand, large bursts do occur and nearly always result in overwhelming a router in the network.

Quick-Start A fundamental aspect of communication in general-purpose, best-effort packet-switched networks is determining an appropriate sending rate. The appropriate sending rate depends on the characteristics of the network path between the two peers (bandwidth, propagation delay, etc.), as well as the amount of load being placed on the network by others at the given time. Traditionally, TCP has used a set of congestion control algorithms for determining this rate. The problem addressed by Quick-Start is how a particular connection on an under-utilized network path can increase its sending rate to take advantage of the available capacity more rapidly than allowed by TCP's traditional congestion control algorithms. Quick-Start is a proposed mechanism for end nodes to request permission from routers along the path to use a higher sending rate. We note that Quick-Start is not in fact a congestion control mechanism, in that it doesn't detect or respond to congestion, and does not replace the traditional congestion control mechanisms of the transport protocol; rather, Quick-Start is an optional mechanisms that flows could use to get approval from routers to send at a high sending rate on a significantly-underutilized path. We are proposing Quick-Start as an Experimental standard in the IETF, and are conducting on-going research about its strengths and weaknesses.

2.3 Internet Routing

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

Towards a Next Generation Inter-domain Routing Protocol: BGP, the current inter-domain routing protocol, is known to suffer from several pressing problems that render the current infrastructure vulnerable to attacks and threaten to impede Internet growth. This project proposes a new inter-domain routing protocol that combines features from both link-state and path-vector routing. This hybrid Link-state Path-vector protocol, called HLP, addresses five specific issues with BGP: lack of scalability, lack of security, poor convergence, lack of fault isolation, and lack of transparency for problem diagnosis. HLP preserves the basic operational and economic model of BGP and only modifies the way in which routing information is propagated. Our hope is that this proposal will stimulate a new debate about the nature of a next-generation BGP.

Reduced State Routing: In today's Internet core, routers store forwarding state proportional to the number of edge networks. As the Internet grows and core line rates increase, routers require memories that are increasingly fast and large?, and are correspondingly increasingly expensive and difficult to engineer. This project proposes Reduced-State Routing (RSR), in which core routers require state only concerning the network topology within a two-hop radius, and thus of a size independent of the total number of Internet edge networks. RSR achieves this feat by routing geographically using two sets of node addresses: virtual coordinates, that are assigned to reflect the link costs within an autonomous system; and geographic coordinates, that correspond to nodes' physical locations. RSR routes greedily on virtual coordinates, and falls back to face routing on geographic coordinates when greedy progress is impossible on virtual coordinates. Unlike previous geographic routing schemes, RSR works on Internet-like graphs (rather than only on wireless-like graphs), and supports policy routing. By simulating RSR on real tier-1 ISP topologies, we can demonstrate that RSR achieves low path stretch, comparable to that caused by policy routing in today's Internet.

2.4 Novel Internet Architectures

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

Delegation-Oriented Architecture: Intermediate network elements, such as network address translators (NATs), firewalls, and transparent caches are now commonplace. The usual reaction in the network architecture community to these so-called middleboxes is a combination of scorn (because they violate important architectural principles) and dismay (because these violations make the Internet less flexible). While we acknowledge these concerns, we also recognize that middleboxes have become an Internet fact of life for important reasons. To retain their functions while eliminating their dangerous side-effects, we propose an extension to the Internet architecture, called the Delegation-Oriented Architecture (DOA), that not only allows, but also facilitates, the deployment of middleboxes. DOA involves two relatively modest changes to the current architecture: (a) a set of references that are carried in packets and serve as persistent host identifiers and (b) a way to resolve these references to delegates chosen by the referenced host.

Packet Obituaries: The Internet is transparent to success but opaque to failure. This veil of ignorance prevents ISPs from detecting failures by peering partners, and hosts from intelligently adapting their routes to adverse network conditions. To rectify this, we propose an accountability framework that would tell hosts where their packets have died. We have a preliminary version of this framework and have analyzed its viability.

Is the Internet Evolvable?: There is widespread agreement on the need for architectural change in the Internet, but very few believe that current ISPs will ever effect such changes. In this project we ask what makes an architecture evolvable, by which we mean capable of gradual change led by the incumbent providers. This involves both technical and economic issues, since ISPs have to be able, and incented, to offer new architectures. Our analysis suggests that, with very minor modifications, the current Internet architecture could be evolvable.

The Internet Impasse: The current Internet is at an impasse because new architectures cannot be deployed, or even adequately evaluated. This project tries to overcome this impasse using virtualization. We propose the use of an shared "virtual testbed" that will enable low-cost experimentation with live traffic and might point the way to easier deployment. This project raises questions about the nature of architecture and the difference between purists and pluralists.

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

2.5 Measurements and Modeling

Sound Internet measurement Conducting an Internet measurement study in a sound fashion can be much more difficult than it might first appear. In this effort we endeavored to formulate a number of strategies drawn from experiences for avoiding or overcoming some of the associated pitfalls. Some of these in particular included dealing with errors and inaccuracies, the importance of associating *meta-data* with measurements, the technique of calibrating measurements by examining outliers and testing for consistencies, difficulties that arise with large-scale measurements, the utility of developing a discipline for reliably reproducing analysis results, and issues with making datasets publicly available.

Reactive measurement: Reactive measurement (REM) is a measurement technique in which one measurement's results are used to decide what (if any) additional measurements are required to further understand some observed phenomenon. While reactive measurement has been used on occasion in measurement studies, what has been lacking is *(i)* an examination of its general power, and *(ii)* a generic framework for facilitating fluid use of this approach. We believe that by enabling the coupling of disparate measurement tools, REM holds great promise for assisting researchers and operators in determining the root causes of network problems and enabling measurement targeted for specific conditions. This project aims to explores REM's power by developing a prototype REM system and applying it to perform a number of measurement studies.

NIMI/SAMI: The NIMI (National Internet Measurement Infrastructure) project focuses on developing and deploying a system for facilitating coordinated measurement from a number of points around the Internet. The final efforts in this project have been towards SAMI (Secure and Accountable Measurement Infrastructure), a major revision of the NIMI architecture that aims to refine its authorization, security, and resource control mechanisms. SAMI is now close to final release.

Modeling enterprise traffic: The characteristics of network traffic *within* an enterprise have gone unexamined in the literature for more than a decade. This project aims to develop such a characterization for modern Internet traffic, as recorded internal to the Lawrence Berkeley National Laboratory. Such basic questions as "What are the dominant types of traffic" and "How do the traffic patterns differ form wide-area Internet traffic" remain unanswered. Thus, this effort has the potential to yield many interesting and possibly surprising insights.

Measuring the evolution of transport protocols: While the Internet's architecture, protocols and applications are constantly evolving, there is often *competing evolution* between various network entities. This competing evolution can impact performance and robustness, and even halt communications in some cases. In our research we have investigated the evolution of TCP the Internet's most heavily used transport protocol, in the context of ongoing changes to the Internet's basic architecture. In particular, we have studied the ways in which so-called "middleboxes" (firewalls, NATs, proxies, etc.) — which change the Internet's basic end-to-end principle — impact TCP. We have explored unexpected interactions between layers and ways in which the Internet differs from its textbook description, including the difficulties various real-world "gotchas" impose on the evolution of TCP (and end-to-end protocols in general). The measurements in our studies have also serves as lessons for efforts to further evolve end-to-end protocols and the Internet architecture.

2.6 Distributed Hash Tables

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

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

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

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

Peering Peer-to-Peer Providers: The early peer-to-peer applications eschewed commercial arrangements and instead established a grass-roots model in which the collection of end-users provided their own distributed computational infrastructure. While this cooperative end-user approach works well in many application settings, it does not provide a sufficiently stable platform for certain peer-to-peer applications (e.g., DHTs as a building block for network services). Assuming such a stable platform isnt freely provided by a benefactor (such as NSF), we must ask whether DHTs could be deployed in a competitive commercial environment. The key issue is whether a multiplicity of DHT services can coordinate to provide a single coherent DHT service, much the way ISPs peer to provide a completely connected Internet. This project has developed various approaches for DHT peering and has analyzed some of the related performance and incentive issues.

CiteSeer on a DHT: As the academic world moves away from physical journals and proceedings to online document repositories, the ability to efficiently locate work of interest among the vast sea of newly-generated papers will become increasingly important. Towards this end, this project has developed SmartSeer, a system that allows users to register personalized continuous queries over the CiteSeer database of technical documents. These users will then be alerted whenever papers that match their queries are put online. SmartSeer has two main design requirements: it should support rich continuous queries (as opposed to simple keyword searches) to allow effective information retrieval and it should be capable of running on a loosely maintained group of unreliable machines donated by multiple organizations (as opposed to assuming a reliable and tightly coupled distributed system). Existing work on distributed continuous query systems fails at least one of these requirements. Our design for SmartSeer is based on Distributed Hash Tables (DHTs), and thereby leverages previous work on DHT-based query systems. A prototype of Smartseer has been implemented and evaluated, and we hope to soon have a publicly available service deployed on Planetlab. Though we evaluate our design only for the SmartSeer application, we believe it also provides useful insights into other distributed and rich continuous query systems (web alerts, news alerts etc).

2.7 Security, Malware, and Intrusion Detection

Enhanced "network telescope" analysis: Network "telescopes", which record packets sent to unused blocks of Internet address space, have emerged as an important tool for observing Internet-scale events such as the spread of worms and the backscatter from flooding attacks that use spoofed source addresses. Previous telescope analyses have produced detailed tabulations of packet rates, victim population, and evolution over time. While such cataloging is a crucial first step in studying the telescope observations, incorporating an understanding of the underlying processes generating the observations allows us to construct detailed inferences about the broader "universe" in which the Internet-scale activity occurs, greatly enriching and deepening the analysis in the process. In this project we applied such an analysis to the propagation of the *Witty* worm, a malicious and well-engineered worm that when released in March 2004 infected more than 12,000 hosts worldwide in 75 minutes. We found that by carefully exploiting the structure of the worm, especially its pseudo-random number generation, from limited and imperfect telescope data we could with high fidelity: extract the individual rate at which each infectee injected packets into the network *prior* to loss; correct distortions in the telescope data due to the worm's volume overwhelming the monitor; reveal the worm's inability to fully reach all of its potential victims; determine the number of disks attached to each infected machine; compute when each infectee was last booted, to sub-second accuracy; identify the effects of NAT on the observations; explore the "who infected whom" infection tree; uncover that the worm specifically targeted hosts at a US military base; and pinpoint *Patient Zero*, the initial point of infection, i.e., the IP address of the system the attacker used to unleash Witty.

Hardware support for network intrusion detection: With ever-increasing network speeds and traffic volumes, there is a growing need to support network intrusion detection using custom hardware. Yet designing such hardware in a fashion that is both robust and sufficiently flexible takes great care. Our recent efforts to address this need have been in two different areas, *shunting* and *stream reassembly.*

The *shunting* project aims to develop an architecture that combines the power of high speed network elements with the flexibility of highly programmable network intrusion detection systems (NIDSs). The core of the architecture is a network forwarding element (the "shunt") that works in conjunction with a NIDS by diverting a subset of the traffic stream through the NIDS. Because the NIDS receives the actual traffic itself rather than a copy, the architecture enables the NIDS to instantly block attack traffic (i.e., "intrusion prevention"). The key insight leveraged by architecture is that in many environments, the vast majority of the high-volume traffic is confined to a small fraction of the connections. Furthermore, these high-volume connections are generally of little interest from an intrusion-detection perspective *after* they have been initially established. That is, it is important to analyze the connections' surrounding context (control session, initial authentication dialog, concurrent logins, etc.), but, once established, the connections themselves can be safely skipped. The core of the hardware support is based on extending the usual packet-filter model with lookup tables. Such tables can be indexed at a variety of granularities: individual connections (source address/source port/destination address/destination port tuples); specific source or destination addresses or pairs; and source/destination prefixes. The key is allocating sufficient memory in the network element so that these tables (particularly the first, per-connection) can be large. The element then looks up incoming packets in the given tables to find if they match flavors of traffic specified by the tables. If so, the element executes the action associated with the table element, where the actions are one of: pass-through, shunt, or drop. Such tables allow the NIDS to communicate fine-grained go/no-go/inspect decisions to the network element in a concise manner: it simply sends over new table entries and their associated actions as it makes decisions concerning whether a given connection, pair of hosts, source, or destination is deemed trustworthy, malicious, or undecided. By basing the project's first prototype on 1 Gbps FPGA forwarding elements, we target enabling the architecture to scale to 10 Gbps in the near-term as next-generation FPGAs become available, and then to 40 Gbps based on striping the traffic across four 10 Gbps elements.

The *stream reassembly* project investigates the design of a hardware module to provide the basic operation of correctly reassembling any out-of-sequence packets delivered by an underlying unreliable network protocol such as IP, a task inherent to higher level analysis of traffic carried with a stateful transport protocol such as TCP. This module has applications to hardware support for a range of high-speed network devices performing packet processing at semantic levels above the network layer, such as layer-7 switches, content inspection and transformation systems, and network intrusion detection/prevention systems. This seemingly prosaic task of reassembling the byte stream becomes an order of magnitude more difficult to soundly execute when conducted in the presence of an *adversary* whose goal is to either subvert the higher-level analysis or impede the operation of legitimate traffic sharing the same network path. In this effort, we designed a hardware-based high-speed TCP reassembly mechanism robust against such attacks. Using trace-driven analysis of out-of-sequence packets, we characterized the dynamics of benign TCP traffic and found we can leverage the results to design a reassembly mechanism that is efficient when dealing with non-attack traffic. We then refined the mechanism to keep the system effective in the presence of adversaries. We found that although the damage caused by an adversary cannot be completely eliminated, it is possible to mitigate the damage to a great extent by careful design and resource allocation, and we devised analytic expressions ("Zombie equations") to quantify the trade-off between resource availability and damage from an adversary in terms of the number of compromised machines ("zombies") an attacker must have under their control in order to exceed a specified notion of "acceptable collateral damage."

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

Internet "background radiation": Monitoring any portion of the Internet address space reveals incessant activity. This holds even when monitoring traffic sent to unused addresses, which we term "background radiation." Background radiation reflects fundamentally nonproductive traffic, either malicious (flooding backscatter, scans for vulnerabilities, worms) or benign (misconfigurations). In this collaborative project we developed the first systematic, broad characterization of the Internet's current background radiation. We based our characterizations on data collected from four unused networks in the Internet. Two key elements of our methodology were *(i)* the use of *filtering* to reduce load on the measurement system, and *(ii)* the use of *active responders* to elicit further activity from scanners in order to differentiate different types of background radiation. We analyzed the components of background radiation by protocol, application, and specific exploits, assessed temporal patterns and correlated activity, and investigated variations across different networks and over time. We found a menagerie of activity, with probes from worms and "autorooters" heavily dominating.

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

Investigating "scaledown" for analyzing large-scale Internet phenomena: A major challenge when attempting to analyze and model large-scale Internet phenomena such as the dynamics of global worm propagation is finding appropriate abstractions that allow us to tractably grapple with size of the artifact while still capturing its most salient properties. In this project we investigated "scaledown" techniques for approximating global Internet worm dynamics by shrinking the effective size of the network under study. We explored scaledown in the context of both simulation and analysis, using as a calibration touchstone an attempt to reproduce the empirically observed behavior of the Slammer worm, which exhibited a peculiar decline in average per-worm scanning rate not seen in other worms (except for the later Witty worm, which exhibited similar propagation dynamics). We developed a series of abstract models approximating Slammer's Internet propagation and demonstrated that such modeling appears to require incorporating both heterogeneous clustering of infectibles and heterogeneous access-link bandwidths connecting those clusters to the Internet core. We demonstrated the viability of scaledown, but also explore two important artifacts it introduces: heightened variability of results, and biasing the worm towards earlier propagation.

Lower bounds on speed of worm propagation: "Flash worms" follow a precomputed spread tree using prior knowledge of all systems vulnerable to the worm's exploit. Our previous work suggested that a flash worm could saturate one million vulnerable hosts on the Internet in under 30 seconds. In this effort, we revisited this problem in the context of single packet UDP worms (such as Slammer and Witty). Using current Internet latency measurements for calibration, we found that with careful design, a flash version of Slammer that used the worm could saturate 95% of one million vulnerable hosts on the Internet in 510 milliseconds. We also investigated a similar design for a TCP-based worm, finding it could 95% saturate in 1.3 seconds. These speeds above are achieved with flat infection trees and packets sent at line rates. Such worms are vulnerable to recently proposed worm containment, but could evade these by using deeper, narrower trees. We explored the resilience of such spread trees when the list of vulnerable addresses is inaccurate and showed that modified, robust designs could still attain extremely fast spread propagation. We also examined the implications of flash worms for containment defenses, finding that such defenses must correlate information from multiple sites in order to detect the worm , but the speed of the worm will defeat this correlation unless a certain fraction of traffic is artificially delayed in case it later proves to be a worm.

Upper bounds on worm damage: In this collaborative effort, we explored the question of to what degree do worms potentially represent a substantial economic threat to the U.S. computing infrastructure. An estimate of how much damage might be caused can greatly aid in evaluating how much to spend on defenses. We constructed a parameterized worst-case analysis based on a simple damage model, combined with our understanding of what an attack could accomplish. Although our estimates are at best approximations, we argue that a plausible worst-case worm could cause $50 billion or more in direct economic damage by attacking widely-used services in Microsoft Windows and carrying a highly destructive payload.

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

As part of this effort, we aimed to address a dearth in the research literature of examinations of the practical difficulties of operating network intrusion detection systems (NIDSs) in large-scale environments, especially the extreme challenges with respect to traffic volume, traffic diversity, and resource management. Our evaluation was based on our extensive operational experience. We identified and explored key factors with respect to resource management and efficient packet processing, highlighting their impact using real-world traces. The insights both helped us gauge the trade-offs of tuning a NIDS and led us to explore several novel ways of reducing resource requirements. These new techniques enabled us to improve the state management considerably, as well as balancing the processing load dynamically.

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

Using TRW as a basis, we then developed *worm containment* algorithms suitable for deployment in high-speed, low-cost network hardware. We show that the approximations used to implement TRW efficiently in hardware preserves its ability to rapidly identify and stop scanning hosts with a very low false-positive rate. We also investigated augmenting this approach by devising mechanisms for *cooperation* that enable multiple containment devices to more effectively detect and respond to an emerging infection. This investigation included an exploration of ways that a worm can attempt to bypass containment techniques in general, and ours in particular.

In addition, we generalized the insights from our investigations to argue that substantial anti-worm defenses will need to be embedded in the local area network, creating "Hard-LANs" designed to detect and respond to worm infections. When compared with conventional network-based intrusion detection systems, Hard-LAN devices will require two orders of magnitude better cost/performance, and at least two orders of magnitude better accuracy, resulting in substantial design challenges.

Building a time machine: Insight into past network traffic can have enormous value, both for forensics when analyzing a problem detected belatedly, and to augment real-time decision-making, both to inform *reactive measurement* (see above) and to give additional pinpoint context to a network intrusion detection system (NIDS). This project aims to develop a network *time machine*, which works by passively bulk-recording as much network traffic as possible. The time machine maintains a ring buffer of recent network traffic that matches a given criteria. This criteria needn't be a simple static filter — the decision of what to capture and for how long could be much richer and incorporate more context. This buffer resides in RAM for fast access, with the decision of what traffic to record in the buffer, and how to filter it (e.g., retaining the first N bytes of each connection), being driven off of a collection of policies describing retention for different types of activity. In addition, recorded traffic migrates from RAM to a given allocation of disk space, which is also managed per a collection of policies that again determine which traffic to migrate, how to filter it, and how to expire it as the disk allocation fills up.

A major next step for this project is to integrate it with a real-time NIDS by providing an API by which the NIDS can query activity seen in the recent past for given connections or hosts. This coupling has the potential to greatly offload the NIDS, allowing it to process only lighter-weight request streams and not response streams, unless it sees a problematic request, in which case it can at that point ask the time machine for a copy of the reply to that particular request.

Detecting triggers: Many automated network attacks have the form in which an initial connection to a host triggers (upon success) a subsequent connection, either inbound from the attacker to test whether a "back door" has been established, or outbound from the victim to signal to the attacker that the exploit succeeded. This project aims to develop statistical anomaly detection techniques for detecting such attacks. We view the problem in abstract terms as attempting to identify pairs of connections that are *causally* related. By grouping connections related to the same application into sessions, we are able to formulate a model based on session-arrival statistics to determine when a new arrival that is separate from an existing session has occurred implausibly soon after a previous arrival. This forms the basis for detecting causalities other than those that arise naturally from expected application patterns. However, a major challenge with this work is determining the full set of such expected application patterns.

Bro advanced development: The *Bro* intrusion detection system has served as the basis for ongoing intrusion detection research since the mid 1990s. One of the strengths of the development of Bro has been its ongoing operational use, first at the Lawrence Berkeley National Laboratory (LBNL), and now in addition at the Technical University of Munich and U.C. Berkeley. The "Bro Lite" project, in collaboration with LBNL, aims to develop a version of Bro amenable to broader, production use at sites that do not necessarily have any Bro expertise. Major elements of this effort include *(i)* enhanced documentation, *(ii)* turn-key operation, *(iii)* augmentation of Bro's "signature matching" facilities, a concept already familiar to many operators, *(iv)* development of a Graphical User Interface for exploring the logs Bro produces, *(v)* production-quality support for Bro in terms of bug-tracking, web presence, and mailing list.

2.8 Sensornets

Practical and Robust Geographic Routing: Geographic routing has been widely hailed as the most promising approach to generally scalable wireless routing when geographic information is available. However, the correctness of all currently proposed geographic routing algorithms relies on idealized assumptions about radios and their resulting connectivity graphs. We used testbed measurements to show that these idealized assumptions are grossly violated by real radios, and that these violations cause persistent failures in geographic routing. Having identified this problem, we then fixed it by proposing the Cross-Link Detection Protocol (CLDP), which enables provably correct geographic routing on *arbitrary* connectivity graphs. We confirmed, in simulation and further testbed measurements, that CLDP is not only correct but practical: it incur low overhead, exhibits low path stretch, and always succeeds in real wireless networks.

Beacon-Vector Routing: This project proposes a practical and scalable technique for point-to-point routing in wireless sensornets when geographic information is not available. This method, called *Beacon Vector Routing* (BVR), assigns coordinates to nodes based on the vector of distances (hop count) to a small set of beacons, and then defines a distance metric on these coordinates. Packets are routed greedily, being forwarded to the next hop that is the closest (according to this beacon vector distance metric) to the destination. This approach has been evaluated through a combination of high-level simulation (to investigate scaling), low-level simulation (to investigate dynamics) and a prototype implementation on motes (as a reality check).

Towards a Sensornet Architecture: Wireless sensor networks have the potential for tremendous societal benefit by enabling new science, better engineering, improved productivity, and enhanced security. Research in this area has progressed dramatically in the past decade. The hardware, particularly radio technology, is improving rapidly, leading to cheaper, faster, smaller, and longer-lasting nodes. Many systems challenges, such robust multihop routing, effective power management, precise time synchronization, and efficient in-network query processing, have been tackled and several complete applications, in which all these components have been integrated into a coherent system, have been deployed and demonstrated, including some at relatively large scale. But the situation in sensornets, while promising, also has problems. The literature presents an alphabet soup of protocols and subsystems that make widely differing assumptions about the rest of the system and how its parts should interact. The extent to which these components can be combined to build usable systems is quite limited. In order to produce running systems, various research groups have produced "vertically integrated" designs in which their own set of components are specifically designed to work together, but are unable to interoperate with components from other groups. This greatly reduces the synergy between research efforts, and has impeded progress in the field. We believe that the primary factor currently limiting progress in sensornets is not any specific technical challenge (though many remain, and deserve much further study) but is instead the lack of an overall sensor network architecture. Such an architecture would identify the essential components and their conceptual relationships so that it would become possible to compose components in a manner that transcends particular generations of technology, allows innovation, and promotes interoperability.

2.9 Internet Community Activities

ICSI researchers are highly active in the Internet research community. In addition to the normal academic duties of serving on program committees and editorial boards, ICSI researchers devote substantial time to more practical duties associated with the Internet Engineering Task Force (IETF), the Internet Research Task Force (IRTF), and other organizational activities within the research community. In particular, in 2004 Vern Paxson served as Chair of the IRTF and contributed to a DARPA ISAT study on "Bolt-on Security." Sally Floyd as a member of the Internet Architecture Board (IAB), the technical advisory board to the Internet Society and the IETF. Mark Allman chairs the IRTF's Internet Measurement Research Group (IMRG), co-chairs the IETF's TCP Maintenance and Minor Extensions (TCPM) and Improved Cross-Area Review (ICAR) Working Groups and is a member of the IETF's General Area Review Team.

References

Papers

[1] M. Allman and E. Blanton. Notes on Burst Mitigation for Transport Protocols. December 2004, under submission.

[2] A. Archer, J. Feigenbaum, A. Krishnamurthy, R. Sami and S. Shenker. Approximation and Collusion in Multicast Cost Sharing. Games and Economic Behavior 47, pp. 36 71, 2004. Abstract appears in Proceedings of the 2001 ACM Conference on Electronic Commerce.

[3] K. Argyraki, P. Maniatis, D. R. Cheriton and S. Shenker. Providing Packet Obituaries. In the ACM HotNets workshop, San Diego, CA, November 2004.

[4] H. Balakrishnan, K. Lakshminarayanan, S. Ratnasamy, S. Shenker, I. Stoica and M. Walfish. A Layered Naming Architecture for the Internet. SIGCOMM 04.

[5] H. Balakrishnan, S. Shenker and M. Walfish. Peering Peer-to-Peer Providers. 4th International Workshop on Peer-to-Peer Systems (IPTPS '05), Ithaca, NY, February 2005.

[6] H. Chang, R. Govindan, S. Jamin, S. Shenker and W. Willinger. Towards Capturing Representative AS-level Internet Topologies. Computer Networks, Vol. 44, No. 6, pp. 737–755, 2004.

[7] H. Dreger, A. Feldmann, V. Paxson and R. Sommer. Operational Experiences with High-Volume Network Intrusion Detection. Proc. ACM CCS, October 2004.

[8] W. Eddy, S. Ostermann and M. Allman. New Techniques for Making Transport Protocols Robust to Corruption-Based Loss. ACM Computer Communication Review, 34(5), October 2004.

[9] J. Feigenbaum, C. Papadimitriou, R. Sami and S. Shenker. BGP-based Mechanism for Lowest-Cost Routing. To appear in Distributed Computing. Preliminary version appears in Proceedings of the 2002 ACM Symposium on Principles of Distributed Computing.

[10] J. Feigenbaum, R. Sami and S. Shenker. Mechanism Design for Policy Routing. To appear in Distributed Computing. Preliminary version in Proceedings of the 2004 ACM Symposium on Principles of Distributed Computing.

[11] R. Fonseca, S. Ratnasamy, J. Zhao, C. T. Ee, D. Culler, S. Shenker and I. Stoica. Beacon-Vector Routing: Scalable Point-to-Point Routing in Wireless Sensor Networks. To appear in NSDI 2005.

[12] R. Gummadi, N. Kothari, Y.-J. Kim, R. Govindan, B. Karp and S. Shenker. Reduced State Routing in the Internet. Proceedings of Hotnets-III, 2004.

[13] A. Gurtov and S. Floyd. Modeling Wireless Links for Transport Protocols. ACM CCR, 34(2):85-96, April 2004.

[14] J. Jung, V. Paxson, A. Berger and H. Balakrishnan. Fast Portscan Detection Using Sequential Hypothesis Testing. Proc. IEEE Symposium on Security and Privacy, May 2004.

[15] B. Karp, S. Ratnasamy, Sean Rhea and S. Shenker. Spurring the Adoption of DHTs with OpenHash. In Third International Workshop on Peer-to-Peer Systems (IPTPS) 2004.

[16] Y.-J. Kim, R. Govindan, B. Karp and S. Shenker. Geographic Routing Made Practical. Proceedings of the USENIX Symposium on Networked Systems Design and Implementation, May 2005.

[17] R. Krishnan, J. Sterbenz, W. Eddy, C. Partridge and M. Allman. Explicit Transport Error Notification (ETEN) for Error-Prone Wireless and Satellite Networks. Computer Networks, 46(3), October 2004.

[18] A. Medina, M. Allman and S. Floyd. Measuring Interactions Between Transport Protocols and Middleboxes (postscript, PDF). Internet Measurement Conference 2004, August 2004. TBIT web page.

[19] A. Medina, M. Allman and S. Floyd. Measuring Interactions Between Transport Protocols and Middleboxes. ACM SIGCOMM/USENIX Internet Measurement Conference, Taormina, Sicily, Italy, October 2004.

[20] A. Medina, M. Allman and S. Floyd. Measuring the Evolution of Transport Protocols in the Internet. Under submission, December 2004.

[21] A. Medina, M. Allman and S. Floyd. Measuring the Evolution of Transport Protocols in the Internet. December 2004, under submission.

[22] P. Narasimhan, H. Kruse, S. Ostermann and M. Allman. On the Impact of BER on Realistic TCP Traffic in Satellite Networks. Technical Report 04-005, International Computer Science Institute, November 2004.

[23] R. Pang, V. Yegneswaran, P. Barford, V. Paxson and L. Peterson. Characteristics of Internet Background Radiation. Proc. ACM IMC, October 2004.

[24] V. Paxson. Strategies for Sound Internet Measurement. Proc. ACM IMC, October 2004.

[25] L. Peterson, S. Shenker and J. Turner. Overcoming the Internet Impasse Through Virtualization. Proceedings of the Third Workshop on Hot Topics in Networking (HotNets-III) San Diego, CA, November 2004.

[26] S. Ramabhadran, S. Ratnasamy, J. M. Hellerstein and S. Shenker. Brief Announcement: Prefix Hash Tree. In Proceedings of ACM PODC, St. Johns, Canada, July 2004. (Earlier versions: IRB Tech Report, Feb. 2004 and, IRB-TR-03-011, Jun. 25, 2003).

[27] R. Sommer and V. Paxson. Exploiting Independent State For Network Intrusion Detection. Technical Report TUM-I0420, Technische Universität München, November 2004.

[28] S. Staniford, D. Moore, V. Paxson and N. Weaver. The Top Speed of Flash Worms. Proc. ACM CCS WORM, October 2004.

[29] J. Stribling, I. G. Councill, J. Li, M. Frans Kaashoek, D. R. Karger, R. Morris and S. Shenker. OverCite: A Cooperative Digital Research Library. 4th International Workshop on Peer-to-Peer Systems (IPTPS '05), Ithaca, NY, February 2005.

[30] L. Subramanian, V. Roth, I. Stoica, S. Shenker and R. Katz. Listen and Whisper: Security Mechanisms for BGP. NSDI 2004.

[31] M. Walfish, H. Balakrishnan and S. Shenker. Untangling the Web from DNS. NSDI 2004.

[32] M. Walfish, J. Stribline, M. Krohn, H. Balakrishnan, R. Morris and S. Shenker. Middleboxes No Longer Considered Harmful. Proceedings of USENIX OSDI, San Francisco, CA, December 2004.

[33] N. Weaver, D. Ellis, S. Staniford and V. Paxson. Worms vs. Perimeters: The Case for Hard-LANs. Proc. Hot Interconnects 12, August 2004.

[34] N. Weaver, I. Hamadeh, G. Kesidis and V. Paxson. Preliminary Results Using Scale-Down to Explore Worm Dynamics. Proc. ACM CCS WORM, October 2004.

[35] N. Weaver and V. Paxson. A Worst-Case Worm. Proc. Third Annual Workshop on Economics and Information Security (WEIS04), May 2004.

[36] N. Weaver, S. Staniford and V. Paxson. Very Fast Containment of Scanning Worms. Proc. USENIX Security Symposium, August 2004.

Internet Drafts

[37] M. Allman and J. Kempf. Using Working Group Review Committees. Internet-Draft draft-allman-icar-wg-revcomm-00.txt, April 2004 (work in progress).

[38] S. Floyd and E. Kohler. TCP Friendly Rate Control (TFRC) for Voice: VoIP Variant and Faster Restart. Internet draft draft-ietf-dccp-tfrc-voip-00.txt, November 2004.

[39] S. Floyd and E. Kohler. Profile for DCCP Congestion Control ID 2:TCP-like Congestion Control. Internet draft draft-ietf-dccp-ccid2-08.txt, November 2004 (work in progress).

[40] A. Jain, S. Floyd, M. Allman and P. Sarolahti. Quick-Start for TCP and IP. Internet-Draft draft-amit-quick-start-03.txt, September 2004 (work in progress).

[41] E. Kohler, M. Handley, S. Floyd and J. Padhye. Datagram Control Protocol (DCCP). Internet draft draft-ietf-dccp-spec-09.txt, November 2004 (work in progress).

[42] J. Padhye, S. Floyd and E. Kohler. Profile for DCCP Congestion Control ID 3:TFRC Congestion Control. Internet draft draft-ietf-dccp-ccid3-09.txt, November 2004 (work in progress).

RFCs

[43] R. Atkinson and S. Floyd. IAB Concerns and Recommendations Regarding Internet Research and Evolution. RFC 3869, August 2004.

[44] E. Blanton and M. Allman. Using TCP Duplicate Selective Acknowledgement (DSACKs) and Stream Control Transmission Protocol (SCTP) Duplicate Transmission Sequence Numbers (TSNs) to Detect Spurious Retransmissions. RFC 3708, February 2004.

[45] S. Floyd. Limited Slow-Start for TCP with Large Congestion Windows. RFC 3742, March 2004.

[46] S. Floyd and J. Kempf. IAB Concerns Regarding Congestion Control for Voice Traffic in the Internet. RFC 3714, March 2004.

[47] S. Floyd, T. Henderson and A. Gurtov. The NewReno Modification to TCP's Fast Recovery Algorithm. RFC 3782, April 2004.

3 Algorithms

3.1 Introduction

During 2004 the participants in the Algorithms Group in Berkeley included postdoctoral researchers Jens Gramm, Till Nierhoff, Roded Sharan and Till Tantau, graduate students Mani Narayanan and Eric Xing, undergraduate student Jacob Scott and group leader Richard Karp. A branch of the Algorithms Group in Portland, Oregon was led by John Moody and included Yufeng Liu (postdoc), Matthew Saffell (PhD candidate) and Aron Rempel (research programmer).

Computational biology was the major focus of the algorithms research, with particular emphasis on analysis of genetic regulatory networks and genetic variation and haplotyping. Other significant activities included: learning, games and computational finance; bandwidth allocation and Internet routing; and computational complexity.

3.2 Highlights of the Research

- Within the human genome there are millions of sites called single-nucleotide polymorphisms (SNPs) at which two different nucleotides commonly occur. Genetic variation at these polymorphic sites is linked to many diseases and other phenotypes. *Haplotyping* is an essential computational step in determining, from experimental measurements, the content of selected SNPs for each individual in a population. The haplotyping program Hap, developed by Eran Halperin and Eliezer Eskin while Halperin was a postdoctoral researcher at ICSI in 2002, has been widely adopted by the genetics and medical research communities. Halperin has recently returned to ICSI as a Research Scientist in the Algorithms group.

- Roded Sharan and Richard Karp collaborated with colleagues Trey Ideker (UCSD), Peter Uetz (Stuttgart University) and their research groups in the comparative analysis of protein-protein interaction data from three species: yeast, worm and fly. They developed a widely applicable computational method which provided strong statistical evidence for hundreds of protein complexes and pathways, and thousands of protein functions and protein-protein interactions, that had not previously been observed. Their collaborators provided experimental verification of a selected set of the predicted protein-protein interactions.

- John Moody's group continued its development of Stochastic Direct Reinforcement algorithms, which show promise of being a superior alternative to traditional reinforcement learning methods for solving real world applications.

3.3 Computational Biology

Analysis of Protein-Protein Interaction Networks To elucidate cellular machinery on a global scale, we performed a multiple comparison of the newly-available protein-protein interaction networks of C. elegans, D. melanogaster and S. cerevisiae [31]. This comparison integrates protein interaction and sequence information to reveal 71 network regions that are conserved across all three species and many exclusive to the metazoans. Using this conservation we found statistically-significant support for

4,645 new protein functions and 2,609 new protein interactions. We tested sixty interaction predictions for yeast by two-hybrid analysis, confirming approximately half of these. Significantly, many of the predicted functions and interactions would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what is gleaned from the genome.

The interpretation of large-scale protein network data depends on our ability to identify significant sub-structures in the data, a computationally intensive task. We adapt and extend efficient techniques for finding paths in graphs to the problem of identifying pathways in protein interaction networks. We present linear-time algorithms for finding paths in networks under several biologically-motivated constraints [18, 24]. We apply our methodology to search for protein pathways in the yeast protein-protein interaction network. We demonstrate that our algorithm is capable of reconstructing known signaling pathways and identifying functionally enriched paths in an unsupervised manner. The algorithm is very efficient, computing optimal paths of length 8 within minutes and paths of length 10 in less than two hours.

Pathway Reconstruction The following problems arise in the analysis of biological networks: We have a boolean function of n variables, each of which has some default value. An *experiment* fixes the values of any subset of the variables, the remaining variables assume their default values, and the function value is the result of the experiment. How many experiments are needed to determine (reconstruct) the function? How many experiments that involve fixing at most q values are needed? What are the answers to these questions when an unknown subset of the variables are actually involved in the function? In the biological context, the variables are genes and the values are gene expression intensities. An experiment measures the gene levels under conditions that perturb the values of a subset of the genes. The goal is to reconstruct the particular logic (regulation function) by which a subset of the genes together regulate one target gene, using a small number of experiments, each involving only minor perturbations. We studythese questions under the assumption that all functions belong to a biologically motivated set of so-called chain functions. We gie optimal reconstruction schemes for several scenarios and show their application in reconstructing the regulation of galactose utilization in yeast [6, 7].

Genome Variation and Haplotyping The problem of inferring haplotypes from genotypes of single nucleotide polymorphisms (SNPs) is essential for the understanding of genetic variation within and among populations, with important applications to the genetic analysis of disease propensities and other complex traits. The problem can be formulated as a mixture model, where the mixture components correspond to the pool of haplotypes in the population. The size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. Thus methods for fitting the genotype mixture must crucially address the problem of estimating a mixture with an unknown number of mixture components. We present a Bayesian approach to this problem based on a non-parametric prior known as the Dirichlet process [40]. It also incorporates a likelihood that captures statistical errors in the haplotype/genotype relationship, trading off these errors against the size of the pool of haplotypes. We describe an algorithm based on Markov chain Monte Carlo for posterior inference in our model. The overall result is a flexible Bayesian method that is reminiscent of parsimony methods in its preference for small haplotype pools. We apply our approach to the analysis of both simulated and real genotype data, and compare to extant methods.

The dissection of complex diseases is one of the greatest challenges of human genetics with important clinical and scientific applications. Traditionally, associations were sought between single genetic markers and disease. The availability of large scale SNP data makes it possible for the first time to study the predictive power of genotypes and haplotypes with respect to phenotype data. We present a novel method for predicting phenotype information from genotype data. The method is based on a support vector machine that uses new kernel functions for the similarity between genotypes or their underlying haplotypes. We demonstrate our approach on SNP data for the apolipoprotein gene cluster in baboons, predicting plasma lipid levels with significant success rates, and identifying associations that were not detected using extant approaches.

Computational methods for inferring haplotype information from genotype data are used in studying the association between genomic variation and medical condition. Recently, Gusfield proposed a haplotype inference method that is based on perfect phylogeny principles. A fundamental problem arises when one tries to apply this approach in the presence of missing genotype data, which is common in practice. We show that the resulting theoretical problem is NP-hard even in very restricted cases. To cope with missing data, we introduce a variant of haplotyping via perfect phylogeny in which a *path* phylogeny is sought. Searching for perfect path phylogenies [10]. is strongly motivated by the characteristics of human genotype data: 70% of real instances that admit a perfect phylogeny also admit a perfect path phylogeny. Our main result is a fixed-parameter algorithm for haplotyping with missing data via perfect path phylogenies. We also present a linear-time algorithm for the problem on complete data.

Association studies in populations relate genomic variation among individuals with medical condition. Key to these studies is the development of efficient and affordable genotyping techniques. Generic genotyping assays are independent of the target SNPs and offer great flexibility in the genotyping process. Efficient use of such assays calls for identifying sets of SNPs that can be interrogated in parallel under constraints imposed by the genotyping technology. We study problems arising in the design of genotyping experiments using generic assays [30]. Our problem formulation deals with two main factors that affect the genotyping cost: The number of assays used, and the number of PCR reactions required for sample preparation. We prove that the resulting computational problems are hard, but provide approximate and heuristic solutions to these problems. Our algorithmic approach is based on recasting the multiplexing problems as partitioning and packing problems on a bipartite graph. We tested our algorithmic approaches on an extensive collection of synthetic data and on data that was simulated using real SNP sequences. Our results show that the algorithms achieve near-optimal designs in many cases, and demonstrate the applicability of generic assays to SNP genotyping.

Transcriptional Regulation: Many signals in biological sequences are based on the presence or absence of base signals and their spatial combinations. One of the best known examples in this regard is the signal identifying a core promoter—the site at which the basal transcription machinery starts the transcription of a gene. Our goal is a fully automatic pattern recognition system for a family of sequences that simultaneously discovers the base signals, their spatial relationships and a classifier based upon them. We present a general method for characterizing a set of sequences by their recurrent motifs [25, 26]. Our approach relies on novel probabilistic models for DNA binding sites and modules of binding sites, on algorithms to learn them from data, and on a support vector machine that uses the learned models to classify a set of sequences. We demonstrate the applicability of our approach to diverse instances, ranging from families of promoter sequences to a data set of intronic sequences flanking alternatively spliced exons. On a core promoter data set our results are comparable to the state-of-the-art McPromoter. On a data set of alternatively spliced exons we outperform a previous approach. We also achieve high success rates in recognizing cell cycle regulated genes. These results demonstrate that a fully automatic pattern recognition algorithm can meet or exceed the performance of hand-crafted approaches.

In [42] we address the problem of modeling generic features of structurally but not textually related DNA motifs, that is, motifs whose consensus sequences are entirely different but nevertheless share "metasequence features" reflecting similarities in the DNA-binding domains of their associated protein recognizers. We present MOtifPrototyper, a profile Bayesian model that can capture structural properties typical of particular families of motifs. Each family corresponds to transcriptional regulatory proteins with similar types of structural signatures in their DNA-binding domains. We show how to train MotifPrototypers from biologically identified motifs categorized according to the TRASFAC categorization of transcription factors and present empirical results of motif classification, motif parameter estimation and *de novo* motif detection by using then learned profile models.

We have combined genome-wide transcription factor binding and expression profiling to assemble a regulatory network controlling the myogenic differentiation program in mammalian cells [4]. We identified a cadre of overlapping and distinct targets of the key myogenic regulatory factors (MRFs), MyoD and myogenin, and MEF2. We discovered that MRFs and MEF2 regulate a remarkably extensive array of transcription factor genes that propagate and amplify the signals initiated by MRFs. We found that MRFs play an unexpectedly wide-ranging role in directing the assembly and usage of the neuromuscular junction. Interestingly, these factors also prepare myoblasts to respond to diverse types of stress. Computational analyses identified novel combinations of factors that, depending on the differentiation state, collaborate with MRFs. Our studies suggest unanticipated biological insights into muscle development and highlight new directions for further studies of genes involved in muscle repair and responses to stress and damage.

Analysis of Protein Contact Maps Structure analysis is an important area in computational biology as the function of molecules is highly associated with their structure. Combinatorial models have been developed to capture the structure of molecules that consist of chains of single units, e.g., *contact maps* in the area of protein structure analysis or *arc annotations* in RNA structure analysis. More precisely, a contact map consists of an ordered set of elements $S = \{1, 2, \ldots, n\}$, corrsponding to the sequence of amino acids in a protein, and a set A of pairs (i_1, i_2), $i_1, i_2 \in S$ and $i_1 < i_2$, which are called *arcs* and correspond to the bonds between pairs of amino acids. Nowadays, experimentally determined structural information is available for large numbers of molecules. One way to organize and analyze these data is to classify proteins according to their structure, as is done in publicly available databases like CATH and SCOP.Surprisingly, this classification in many cases still involves human interaction. In a structural approach, the classification considers proteins to belong to the same class if they share structural features, even if their primary sequence may not be similar. Therefore, in the interpretation of data, an important question is to determine whether a known structural pattern occurs in a given molecular structure. This, however, is the straightforward extension of the classical pattern matching problem from strings to contact maps, resulting in the Contact Map Pattern Matching problem. More precisely, given two contact maps $S, A)$ and (S_p, A_p), we ask whether the "pattern" (S_p, A_p) occurs in (S, A).

In general, the pattern matching problem for contact maps is NP-hard. We have determined two large classes of structural patterns for which Contact Map Pattern Matching becomes solvable in polynomial time. In [2] we showed that the problem is solvable in quadratic time if the arc set A_p is nested, and in [8], that the problem is solvable in polynomial time if every two arcs cross. We have found experimentally that, using such structural patterns, we can reliably distinguish protein domains listed in the CATH database as members of a given superfamily from non-members listed as 'structural relatives' of members.

Genome Rearrangements: One of the most promising ways to determine evolutionary distance between two organisms is to compare the order of appearance of orthologous genes in their genomes. The resulting genome rearrangement problem calls for finding a shortest sequence of rearrangement operations that sorts one genome into the other. We provide a 1.5-approximation algorithm for the problem of sorting by transpositions and transreversals, improving on a five-years-old 1.75 ratio for this problem. Our algorithm is also faster than current approaches and requires $O(n^{3/2}\sqrt{\log n})$ time for n genes [13].

3.4 Learning, Games and Computational Finance

During 2004, John Moody's group conducted research on direct reinforcement learning, multi-agent systems, games, time series and computational finance.

Research advances included further development of Stochastic Direct Reinforcement algorithms, which show promise of being a superior alternative to traditional reinforcement learning methods for solving real world applications. Other results were obtained in competitive and cooperative games (in particular the evolution of cooperation), applications of reinforcement learning to computational finance and development of machine learning methods for time series prediction. Liu secured a position at a top tier financial firm in Fall 2004 after completing a year as a postdoc at ICSI, and Saffell is scheduled to complete his PhD thesis under Moody's guidance in Spring 2005.

Moody continued to focus on building a hedge fund business, being only part-time at ICSI (17April issue of Bloomberg Markets in an article entitled "From Harvard to Hedge Funds" and in international broadcasts on Bloomberg Television.

3.5 Bandwidth Allocation and Internet Routing

Fair Bandwidth Allocation A fundamental goal of Internet congestion control is to ensure fairness by selectively dropping packets from flows that are receiving more than their fair share of bandwidth. The most effective known algorithms for detecting and selectively dropping high-rate flows at a router are based on random hashing or random sampling of packets and give only probabilistic guarantees. The known deterministic algorithms either require excessive storage or fail to guarantee fairness. In a simplified theoretical setting we show that the detection and selective dropping of high-rate flows can be accomplished deterministically without maintaining an excessive amount of state [16]. Given an arriving stream of packets, each labeled with the name of its flow, our algorithm drops the minimum number of packets required to guarantee that, in every consecutive subsequence of the output stream, no flow has significantly more than its fair share of the packets. The main results of the paper are tight bounds on the worst-case storage requirement of this algorithm. The combinatorial problems that are solved to produce these bounds may be of independent mathematical interest.

Optimal Flow Distribution In [17] we consider a simple network flow problem in which there are n channels directed from a source to a sink. The channel capacities are unknown and we wish to determine a feasible network flow of value D. Flow problems with unknown capacities arise naturally in numerous applications, including inter-domain traffic routing in the Internet, bandwidth allocation for sending files in peer-to-peer networks, and the distribution of physical goods among different points of sale. We study protocols that probe the network by attempting to send flow of at most D units through the entwork. If the flow is not feasible, the protocol is told on which channels the capacity was exceeded (binary feedback) and pssibly also how many units of flow were successfully sent on these channels (throughput feedback).For the latter, more informative, type of feedback we present optimal protocols for minimizing the number of rounds needed to find a feasible flow and for minimizing the total amount of wasted flow. For binary feedback, we show that one can exploit the fact that the network capacities may be larger than the demand D. We present a protocol for this situation that is asymptotically optimal and finds a solution more quickly than the generalized binary search protocol previously proposed in the literature. For the special case of two channels we present a protocol that is optimal and outperforms binary search.

3.6 Computational Complexity

In [37] we continued previous research on the complexity of reachability problems. Several results on this complexity from the literature can be summed up as: “The smaller the diameter of a graph, the easier it is to find paths between its vertices.” The new research results can be summarized similarly: “The smaller the independence number of a graph, the easier it is to find paths between its vertices.”

References

[1] I. Adler, H.S. Ahn, R.M. Karp and S.M. Ross. A Probabilistic Model for the Survivability of Cells. Submitted, 2004.

[2] J. Alber, J. Gramm, J. Guo and R. Niedermeier. Computing the Similarity of Two Sequences with Nested Arc Annotations. Theoretical Computer Science, 312(2-3): 337-358, 2004.

[3] A. Ben-Dor, T. Hartman, R. M. Karp, B. Schwikowski, R. Sharan and Z. Yakhini. Towards Optimally Multiplexed Applications of Universal Arrays. Journal of Computational Biology, 11(2-3), pages 476–492, 2004.

[4] A. Blais, M. Tsikitis, D. Acosta-Alvear, R. Sharan, Y. Kluger and B. D. Dynlacht. A Gene Expression Network Governing Mammalian Myogenesis. Submitted to Genes and Development.

[5] J. Elson, R.M. Karp, C.H. Papadmitriou and S. Shenker. Global Synchronization in Sensornets. LATIN 2004: 609-624.

[6] I. Gat-Viks, R. Shamir, R. M. Karp and R. Sharan. Reconstructing Chain Functions in Genetic Networks. Proceedings of the Ninth Pacific Symposium on Biocomputing (PSB'04), pages 498–509, 2004.

[7] I. Gat-Viks, R. M. Karp, R. Shamir and R. Sharan. Reconstructing Chain Funcations in Genetic Networks. Submitted to SIAM Journal on Discrete Mathematics.

[8] J. Gramm. A Polynomial-Time Algorithm for the Matching of Crossing Contact-Map Patterns. Proc. 4th International Workshop on Algorithms in Nioinformatics (WABI2004).

[9] J. Gramm, T. Nierhoff, R. Sharan and T. Tantau. On the complexity of haplotyping via perfect phylogeny. In Proceedings of Second RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotypes, pages 35–46, 2004.

[10] J. Gramm, T. Nierhoff and T. Tantau. Perfect path phylogeny haplotyping with missing data is fixed-parameter tractable. In Proceedings of the 2004 International Workshop on Parameterized and Exact Computation, volume 3162 of Lecture Notes in Computer Science, pages 174–186. Springer-Verlag, 2004.

[11] E. Halperin and R.M. Karp. The Minimum-Entropy Set Cover Problem. ICALP 2004: 733-744.

[12] E. Halperin, R.M. Karp. Perfect Phylogeny and Haplotype Assignment. RECOMB 2004: 10-19.

[13] T. Hartman and R. Sharan. A 1.5-Approximation Algorithm for Sorting by Transpositions and Transreversals. Journal of Computer and System Sciences, in press.

[14] R.M. Karp. Algorithms for Inferring Cis-Regulatory Structures and Protein Interaction Networks. RECOMB 2004: 45.

[15] R.M. Karp. The Role of Experimental Algorithms in Genomics. WEA 2004: 299-300.

[16] R.M. Karp. Guaranteeing Fair Bandwidth Allocation Without Per-Flow State. Submitted.

[17] R.M. Karp, T. Nierhoff and T. Tantau. Optimal Flow Distribution Among Multiple Channels with Unknown Capacities. GRACO 2005 (to appear).

[18] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell and T. Ideker. PathBLAST: A tool for alignment of protein interaction networks. Nucleic Acids Research, 32, pages W83-W88, 2004.

[19] G. Kimmel, R. Sharan and R. Shamir. Computational Problems in Noisy SNP and Haplotype Analysis: Block Scores, Block Identification and Population Stratification. INFORMS Journal on Computing, 16, pages 360–370, 2004.

[20] J. Moody, Y. Liu, M. Saffell and K. Youn. Stochastic Direct Reinforcement: Application to Simple Games with Recurrence. Appears in Artificial Multiagent Learning, Sean Luke et al. editors, AAAI Press, Menlo Park, 2004.

[21] M. Narayanan and R.M. Karp. Gapped Local Similarity Search with Provable Guarantees. WABI 2004: 74-86.

[22] M. Ogihara and T. Tantau. On the reducibility of sets inside NP to sets with low information content. Journal of Computer and System Sciences, 69(4):299–324, 2004.

[23] I. Pe'er, T. Pupko, R. Shamir and R. Sharan. Incomplete Directed Perfect Phylogeny. SIAM Journal on Computing, 33(3), pages 590–607, 2004.

[24] J. Scott, T. Ideker, R.M. Karp and R. Sharan. Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks. To appear in Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB'05).

[25] E. Segal and R. Sharan. A Discriminative Model for Identifying Spatial Cis-Regulatory Modules. Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB'04), pages 282–289, 2004.

[26] E. Segal and R. Sharan. A Discriminative Model for Identifying Spatial Cis-Regulatory Modules. Submitted to Journal of Computational Biology.

[27] R. Shamir and R. Sharan. A Fully Dynamic Algorithm for Modular Decomposition and Recognition of Cographs. Discrete Applied Mathematics, 136, pages 329–340, 2004.

[28] R. Sharan, A. Ben-Hur, G.G. Loots and I. Ovcharenko. CREME: Cis-Regulatory Module Explorer for the Human Genome. Nucleic Acids Research, 32, pages W253-W256, 2004.

[29] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir and R. M. Karp. Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data. Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB'04), pages 141–149, 2004.

[30] R. Sharan, A. Ben-Dor and Z. Yakhini. Multiplexing Schemes for Generic SNP Genotyping Assays. Proceedings of the Ninth Pacific Symposium on Biocomputing (PSB'04), pages 140–151, 2004.

[31] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp and T. Ideker. Conserved Patterns of Protein Interaction in Multiple Species. Proc. Natl. Acad. Sci. USA, in press.

[32] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir and R. M. Karp. Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data. Journal of Computational Biology, in press.

[33] R. Sharan, J. Gramm, Z. Yakhini and A. Ben-Dor. Multiplexing Schemes for Generic SNP Genotyping Assays. Submitted to Journal of Computational Biology.

[34] A. Tanay, R. Sharan, M. Kupiec and R. Shamir. Revealing Modularity and Organization in the Yeast Molecular Network by Integrated Analysis of Highly Heterogeneous Genomewide Data. Proc. Natl. Acad. Sci. USA, 101(9), pages 2981–2986, 2004.

[35] A. Tanay, R. Sharan and R. Shamir. Biclustering Algorithms: A Survey. Submitted to Handbook on Computational Molecular Biology.

[36] T. Tantau. Comparing verboseness for finite automata and Turing machines. Theory of Computing Systems, 37(1):95–109, 2004.

[37] T. Tantau. A logspace approximation scheme for the shortest path problem for graphs with bounded independence number. In Volker Diekert and Michel Habib, editors, Proceedings of the 21st International Symposium on Theoretical Aspects of Computer Science, STACS 2004, volume 2996 of Lecture Notes on Computer Science, pages 326–337. Springer-Verlag, 2004.

[38] T. Tantau. Strahlende Präsentationen in LaTeX. Die TeXnische Kom"odie, 2/2004.

[39] T. Tantau. Über strukturelle Gemeinsamkeiten der Aufzählbarkeitsklassen von Turingmaschinen und endlichen Automaten. In Ausgezeichnete Informatikdissertationen 2003, Lecture Notes in Informatics, pages 189–198. Springer-Verlag, 2004.

[40] E. Xing, R. Sharan and M.I. Jordan. Bayesian Haplotype Inference via the Dirichlet Process. Proceedings of the Twenty-First International Conference on Machine Learning, pages 879–886, 2004.

[41] E.P. Xing, W. Wu, M.I. Jordan and R.M. Karp. LOGOS: A Modular Bayesian Model for *de novo* Motif Detection. J. Bioinformatics Comput Biol. 2, 127-154, 2004.

[42] E.P. Xing and R.M. Karp. MotifPrototyper: A Bayesian Profile Model for Motif Families. PNAS, 2004.

4 Artificial Intelligence and its Applications

The Artificial Intelligence group continues its long term study of language, learning, and connectionist neural modeling. The scientific goal of this effort is to understand how people learn and use language. The applied goal is to develop systems that support human-centered computing through natural language and other intelligent systems. The work of the group is conducted within four articulating efforts. The first is a long-standing collaboration between J. Feldman, S. Narayanan, and G. Lakoff (UCB Linguistics) on the Neural Theory of Language (NTL) project. The second is a project headed by Charles Fillmore to develop on-line lexical resource (FrameNet) to be used in semantically rich natural language tasks. The third is an effort in computational neuroscience primarily headed by S. Narayanan and L. Shastri to build neurally plausible models of memory and cognition. The fourth is an effort headed by Paul Kay and ICSI alumnus Terry Regier (now at U. Chicago) on analyzing language data from the World Color Survey (WCS) to test the theory of color naming universals first proposed by Brent Berlin and Paul Kay. In all these projects, there is continuing close cooperation with other groups at ICSI, at UC Berkeley, and with external sponsors and other partners. This reports summarizes the progress made and accomplishments of the four projects in 2004.

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

The NTL project of the AI group works in collaboration with other units on the UCB campus and elsewhere. It combines basic research in several disciplines with applications to natural language processing systems. Basic efforts include studies in the computational, linguistic, neurobiological and cognitive bases for language and thought and continues to yield a variety of theoretical and practical findings.

One ongoing applied effort (called EDU for Even Deeper Understanding) has been in operation since July 2000, with multi-year funding from the Klaus Tschira Foundation. In past years, Robert Porzel, of EML, joined our group for a year and John Bryant from ICSI spent the last half of 2002 working at EML in Heidelberg. There were several mutual visits in 2004 and a new collaborative effort on grammar learning, based on Nancy Chang's forthcoming thesis, was started. A major aspect of this collaboration has been the international workshops on Scalable Natural Language Understanding Systems (SCANALU), first held in Heidelberg. A second Scanalu workshop was held in Boston in May, 2004 and brought together several groups working on embodied language. Two papers from the NTL group [9, 10] were presented at this workshop.

This endeavor is also closely linked to the SmartKom successor SmartWeb project, which is discussed in the Speech section of this annual report. Another cooperation between the Speech and AI groups is the human interface section of the CITRIS project of the state of California. CITRIS is a large multi-disciplinary effort that has many subprograms, one of which is an effort at ICSI, the Berkeley Center for the Information Society (BCIS). BCIS has been restructured in 2004, but is continuing its efforts; a brief summary is included as part of this report.

Starting in 2002 the NTL group began working on another large cooperative NLU project, after winning in a highly competitive grant competition in the AQUAINT program of the U.S. Defense Department ARDA organization. The NTL group teamed with Prof. Marti Hearst (SIMS, UCB) and Prof. Chris Manning (Stanford) to study deep inferencing techniques and corpus based techniques for deriving the conceptual semantics needed to achieve this. The first phase of this was successfully completed in 2004.

In 2004, the NTL group demonstrated several of its results at various ARDA workshops and meetings. The group was also successful in attaining funding for a highly competitive AQUAINT Phase 2, in collaboration with Stanford and U. Texas at Dallas. Our effort is being integrated into an ambitious overall program to significantly advance the automated analysis of information. The new AQUAINT project with Narayanan as the ICSI Prinicpal Investigator makes significant use of our basic work on both grammar and inference and is also contributing to it. Section 4.1.3 describes the use of NTL inference models for Question Answering. The new effort also involves significant funding for the FrameNet project, described in Section 4.2.

One core NTL computational question is finding the best match of constructions to an utterance in linguistic and conceptual context. The task of finding a best-fit analysis and approximate answers (answers are often not exact or correct) presents a more tractable domain than exact symbolic matching. More importantly, our integrated constructions are decidedly not purely syntactic nor context-free. We believe that incorporating constraints both from semantics and context is essential to sufficiently constrain the solution space and make it possible in practice to build best-fit construction matchers of the required scale.

John Bryant, a CS doctoral student, has completed a Masters thesis on this topic and is continuing for his doctorate. His Masters project is now in use at other labs internationally. In 2004, he presented this work and its extensions to an international audience at the Robust Methods in Analysis of Natural language Data, workshop in Geneva, Switzerland [11]. This work in being extended to include discourse and situational context in the semantic best fit computations.

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

We also devised an experimental means by which to test the psychological reality of construal, the variable, context-specific understanding of the semantic pole of linguistic constructions. An initial paper on this was presented at the 2003 Cognitive Science Conference. Several follow on studies have also been published in the 2004 Cognitive Science Conference [13, 8], the Berkeley Linguistics Society meeting and elsewhere.

Nancy Chang and others have continued developing representations and algorithms useful for an embodied approach to language acquisition and use. She worked with colleagues to flesh out different aspects of a simulation-based approach to language understanding, including a formal representation for linguistic constructions. A version of the formalism is incorporated into her thesis research, which focuses on the development of an algorithm that learns such constructions from a set of utterance-situation pairs. The construction learning work was presented at the Stanford Child Language forum [12].

In 2004 there was a very significant increase in the use of the group's results in UCB courses and in linguistics research. Collaboration with the FrameNet project has been broadened and deepened with positive results for both efforts, some of which are described in this report. S. Narayanan and J. Feldman ran an interdisciplinary class in Spring 2004 and several of the research efforts from that class are being incorporated into the project. George Lakoff is using ECG as the basis for his graduate seminar in Spring 2005. Several new UCB doctoral students have become involved with the group including John Bryant, Ellen Dodge, Marc, Etlinger, Olya Gurevich, Eva Mok, Shweta Narayan, and Steven Sinha.

The NTL work is receiving increasing broad attention. Jerome Feldman was a keynote speaker at the 2004 Coling Conference and a related workshop on language understanding in Geneva, Switzerland. He and Srinvas Narayanan also gave an invited tutorial at the Fall 2004 AQUAINT symposium in Baltimore. Narayanan (with S. Harabagiu) also presented a tutorial at the 2004 HLT/NAACL conference entitled "Semantic Structures for Question Answering" [31] in Boston.

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. The first of these is a large continuing effort on modeling Services on the Semantic Web. A second effort is developing multimedia resources on the Semantic Web. A third is deploying previous NTL work on expressive probabilistic inference models in Question Answering systems. A fourth involves applications of FrameNet for Natural Language Processing (described in the FrameNet section (Section 4.2.5)).

4.1.1 Semantic Web Services

The Semantic Web is an exciting vision for the evolution of the World Wide Web. Adding semantics enables structured information to be interpreted unambiguously. Precise interpretation is a necessary prerequisite for automatic Web search, discovery and use. Services are a particularly important component of the Semantic Web. A semantic service description language can enable a qualitative advance in the quality and quantity of e-commerce transactions on the Web. The OWL Services Coalition (Narayanan was a founding member) under the guise of OWL-S, has taken some important first steps in this direction.

The model of actions, processes and events developed within the NTL project provides a natural, distributed operational semantics that may be used for simulation, validation, verification, automated composition and enactment of OWL-S-described Web services. In 2004, OWL-S was formally accepted by W3C as a member submission. http://www.w3.org/Submission/OWL-S has the details.

4.1.2 Multimedia Resources on the Semantic Web

In 2004, Narayanan was a co-PI in the ARDA sponsored Video Event Taxonomy project which was a targeted multi-institutional effort comprising of researchers from the University of Southern California, SRI International, the University of Maryland and ICSI, Berkeley. The project developed a Video Event Representation Language (VERL) to provide a common representational framework and ontology for describing video events. Furthermore, the project developed a common annotation language, VEML, to serve as a uniform data model to facilitate model co-development and interchange of annotated video data.

To enable these efforts to tap into the various tools, ontologies, and resources on the Semantic Web, S. Narayanan and student Matt Gerber implemented VERL and VEML in the Semantic Web markup language, OWL. The ICSI effort provides a uniform API that permits linkage to various ontologies and tools on the Semantic Web while providing programmatic access to the VERL and VEML annotated data for use by researchers on the Semantic Web. We also build automatic translation tools to take existing XML-based video event annotations and convert them to the new OWL format.

The OWL representations of VERL and VEML data have significant advantages over their XML counterparts. The first is that any amount of information from the two projects can be combined into the same graph effectively integrating data from the two projects Groups developing tools that utilize data generated by the projects will not have to worry about the locations, names, and structures of project-specific files because in its OWL-encoded form the data is more self-contained and syntactically homogeneous. This neutrality facilitates the sharing and reuse of the data, and is a key feature of the Semantic Web.

Secondly, defining classes of resources and properties to relate these classes, their attributes, and extensions makes the task of data processing easier. For example, searching the VERL ontology database for all sub-events of a particular event could be an expensive task if the XML files involved had to be processed in their entirety. The OWL representation, combined with efficient tools that are readily available makes the task more feasible.

4.1.3 Semantic Inference for Question Answering

The ability to answer complex questions posed in Natural Language depends on (1) the ability of the analyzer to extract semantic information from language input, (2) the expressiveness and depth of the available semantic representations and (3) the inferential mechanisms they support. In 2004, ICSI, in collaboration with University of Texas, Dallas and Stanford University developed a Question Answering (QA) architecture where questions are analyzed and candidate answers generated by (1) identifying predicate argument structures and semantic frames from the input and (2) performing structured probabilistic inference using the extracted relations in the context of a domain and scenario model [32, 33].

A novel aspect of our system is a scalable and expressive representation of domain and scenario models based on Coordinated Probabilistic Relational Models (CPRM) developed by the NTL group at ICSI [34]. CPRM combine the modular and structured probabilistic framework of Probabilistic Relational Models (PRM) with the flexible control and coordinated dynamics of extended Petri nets. These can be seen as extending an ongoing community effort to add more sophisticated temporal and control capabilities to probabilistic inference models. We have implemented the CPRM model of inference and are currently testing it for use in Question Answering as part of the ARDA sponsored AQUAINT program in collaboration with Stanford University and University of Texas, Dallas. Initial results of applying the CPRM model to Question Answering can be found in the Coling paper by Narayanan and Harabagiu [33].

The current version of our CPRM simulator is installed at the University of Texas where model building and QA experiments are under way to test the utility of the system compared to a state of the art baseline QA system.

We have also started a new collaboration with the Center for Non-Proliferation Studies in Monterey to investigate the utility of CPRM inference techniques for analyzing data, cases and decision processes from their database of terrorist capabilities and incidents. This collaboration also involves FrameNet based semantic processing of CNS documents described in Section 4.2.

4.2 FrameNet, C. Fillmore

4.2.1 Background

The FrameNet project (`http://www.icsi.berkeley.edu/~framenet`), housed at ICSI, funded by NSF (1997-2003)[1] and more recently by DARPA[2], is building a semantically-rich lexicon of English and a set of annotated texts exemplifying more than 600 semantic frames. The theory of Frame Semantics [19] states that each meaning of each word is associated with a semantic **frame** which represents the conceptual structure that underlies it. The frame contains a set of **frame elements**, which are frame-specific names and definitions for the participants and props involved in the situation described by the frame. Most of the effort has gone into building a lexicon of more than 8,000 predicators with detailed information about how their semantic arguments can be expressed in sentences, based on computer-assisted annotation of more than 130,000 example sentences. Applications under study include automatic frame recognition (roughly equivalent to word sense disambiguation), automatic semantic role labeling, machine translation, information extraction, question answering, and text understanding.

The corpus from which the examples were extracted was initially the British National Corpus (100M running words, balanced across genres and domains); we have now added an American Newspaper Corpus from Linguistic Data Consortium (`http://www.ldc.upenn.edu`). We are actively participating in the development of the new American National Corpus, headed by Prof. Nancy Ide of Vassar College, and are using those portions which are currently available.

The software, built in-house, uses client-server model which allows us to run the thin client software on any platform that supports Java, while centralizing the logic in an

[1] Through grants IRI-9618838 "Tools for Lexicon Building" (1997-2000) and HCI-0086132, "FrameNet++: An Online Lexical Semantic Resource and its Application to Speech and Language Understanding" (2000-2003) and a NSF subcontract with D. Jurafsky and M. Palmer for adding FrameNet-style semantic annotations to the PropBank corpus. (2003-2005)

[2] For work on aligning FrameNet data with other lexical and ontological resources.

application server under JBOSS. The application server talks to the MySQL database that contains all the data, as do a set of web-bases report generators. The software can also be distributed to collaborators building similar databases, either for other languages or in specialized domains.

4.2.2 Current Tasks

Continuous Text Annotation

A major new effort of FrameNet during 2004 has been the annotation of continuous texts. Our previous work has been directed toward building a lexicon, using individual sentences as examples of the uses of each lexical unit, regardless of where in the corpus they came from. In general, each sentence was annotated with regard to only one LU.

This year, we have exhaustively annotated continuous texts from two sources, some from the Penn TreeBank project (`http://www.cis.upenn.ecu/~treebank`) and some from the website of the Center for Non-Proliferation Studies (`http://cns.miis.edu/`). This means that we have annotated essentially every argument of each predicator of every sentence of each text. We define predicators not only as verbs (e.g. *to marry*), but also include argument-taking nouns (e.g. *wife*), relational adjectives (e.g. *engaged (to)*), adverbs, etc.; usually, at least two or three words per sentence are annotated. The multi-layered annotation represents a rich semantics of each sentence which should serve (1) as a first stage in a new generation of NLP systems and (2) for comparisons between the FrameNet method of annotation and those of other projects, particularly the Penn TreeBank project and the SALSA project (discussed below).

Core Frame Elements and Data Consistency

A major effort was also made during 2004 to make the existing data more consistent with our evolving theories of Frame Semantics. In the first place, we instituted a three-way distinction among classes of frame elements, dividing the FEs for each frame into **Core**, **Peripheral** and **Extra-thematic**. (Briefly, those that are necessary for the frame itself, those that are ontologically presupposed by the frame, and those that, strictly speaking belong to other frames, but frequently occur along with the frame under study.) We then undertook to check all the existing annotation to ensure that (1) all core FEs were annotated on every sentence in each frame and (2) there is consistency between core FEs and nuclear syntactic positions (such as subject and object of active voice verbs). We have not completed this task, but the vast majority of inconsistencies have been removed.

This clean-up effort was undertaken in anticipation of the data release (R1.2) which is now forthcoming. All of the more than 130,000 annotated sentences were checked automatically for a variety of errors (including inaccurate tags for part-of-speech, etc.) and a substantial portion were also reviewed by the staff. In many cases, inaccurate annotation was removed, so that the total number of annotated sentences in the database is close to what it was at the time of the last data release, but we are confident that the greater accuracy and consistency will produce better results for machine learning systems trained on the newly released data.

Frame-to-frame Relations

Since we now have more than 600 frames, many users of FrameNet data have asked us how they are related to each other and to general thematic roles, such as Agent, Theme/Patient, Instrument, Source, Path, Goal, etc. To explain this, we have developed frame-to-frame relations, locating frames in a hierarchy [20]; at the top level there are very general frames such as Event, State, Intentionally_Act, etc. whose frame elements correspond roughly to these traditional thematic roles. Lower-level frames inherit some of their properties from these general frames. The hierarchy of frame relations includes a similar hierarchy of frame elements, so that the FEs of lower frames are subtypes of the FEs of the top-level frames via a chain of FE-FE bindings. We define three major frame-frame relations, **Inheritance** proper (in which all the core FEs of the parent frame are bound to FEs in the child), **Using**, in which only some of the core parent FEs are bound to child FEs, and **Subframe**, in which the child frames represent stages in a complex event defined by the parent frame. The frame-frame relations were greatly expanded in 2004, currently, numbering 194 frame-frame inheritance relations, 217 using relations, and 80 subframe relations, along with dozens of less prominent relations. We have recently begun work on a web-based viewer for these relations, which should faciliate development work on the frame hierarchy.

Data Releases, Data Users, and Website

To date, there have been three FrameNet data releases, and a fourth (R1.2) is expected in mid-February, 2005. The full set of data is available both in HTML format (for local browsing copies) and XML (for loading into existing NLP systems). In addition, beginning with R1.2, the frame, FE, frame relation and annotation data will be available in OWL, suitable for RDF browsing (.e.g with Protege) and semantic web applications. We hope to release an OWL representation of the lexical entry data soon.

Requests for downloading the FrameNet data arrive daily from individual researchers and from research institutions around the world, providing growing evidence of its usefulness in various areas. More than 300 researchers have downloaded some version of the FrameNet data. Their purposes vary greatly, and are hard to summarize, but one can get a general impression from the following areas which recent downloaders have listed as their research interests, as shown in the table below (Many users indicated more than one interest.):

Interest	No. Users
Natural Language Understanding	87
Semantic Parsing	71
Information Extraction	68
Research in Lexical Semantics	58
Word Sense Disambiguation	62
Question Answering	47
Lexicography	37
Natural Language Generation	29
Machine Translation	27
Teaching Lexical Semantics	22

4.2.3 Collaborative Projects

Semantic Parsing

There has been considerable interest in finding automatic ways of labeling the fillers of semantic argument slots around predicators, since the seminal article [22] (G & J), followed soon after by [21], which used a maximum entropy learning method and [43], which used a HMM model, and learned frame recognition as well. In 2004, a task organized as part of SENSEVAL-3 used FrameNet data as a training corpus for automatic semantic parsing (http://www.clres.com/SensSemRoles.html), essentially replicating the G & J task; 8 teams submitted different types of systems in this competition, some of them obtaining considerably better results than the G & J benchmark.

FrameNets and related projects in other languages

Projects related to FrameNet are underway in several countries:

Spanish FrameNet: Work on Spanish FrameNet has proceeded smoothly in 2004, covering roughly 350 lexical units. SFN, based at Universidad Autónoma de Barcelona, is using the full set of software used in Berkeley, and a completely parallel database. Of course the lexical units are Spanish; the vast majority of the frames have been demonstrated to work as is in Spanish, but some new frames and FEs have been created to handle know differences in, for example, verbs related to motion. (`http://gemini.uab.es/SFN/`) The head of SFN, Prof. Carlos Subirats, has visited ICSI for 4 months in 2004, supervising work in Barcelona via e-mail and telephone and discussing questions as they arrive with staff in Berkeley.

SALSA and German FrameNet: The SALSA project at the Universität des Saarlandes, Saarbrücken. (`http://www.coli.uni-saarland.de/cl/projects/salsa/`) is annotating a German newswire corpus using their own software, but based on FrameNet frames and FEs. They have been collaborating closely with the ICSI group on questions of crosslinguistic frame differences, by telephone, e-mail and site visits during 2004. Prof. Hans C. Boas of U Texas at Austin is seeking support to build a German FrameNet based on the work of SALSA.

Japanese FrameNet: Work on Japanese is at an early stage; a corpus and corpus search tools have been created, and several papers have been published on foreseeable differences in lexicalization between Japanese and English, in terms of frame semantics.

Other languages: BiFrameNet (`http://www.cs.ust.hk/~hltc/Biframenet`, [14]) is a new attempt to build a Chinese FrameNet, including annotated example sentences, entirely automatically, using an existing Chinese ontology, HowNet. Various researchers have also expressed interest in building lexica based on FrameNet for other languages, among them French (exploratory proposal submitted), Rumanian, Bulgarian, Swedish, Hindi, and Hebrew. The first meeting of the Crosslingual FrameNet group took place at ICSI in October, 2004, attended by representatives from ICSI, Saarbrücken, Barcelona, Tokyo, Austin, and Sweden. The members compared research at each site, discussed crosslingual differences in frames and LUs and made plans for ongoing cooperation.

Center for Non-Proliferation Studies

In September, 2004, FrameNet staff and others from ICSI met with staff from the Center for Non-Proliferation Studies to discuss the possibility of cooperation in semantic annotation of texts in the domains of terrorism and proliferation of weapons of mass destruction. FrameNet has annotated several texts from their website, and we are now working on obtaining a substantial set of data from them in an appropriate format for importation and annotation. We are hopeful that some funding can be obtained to facilitate collaboration between the two centers.

FrameSQL

A web-based browser for FrameNet data has been built independently in Japan by Professor Hiroaki Sato of Senshu University, called FrameSQL. In 2004, Prof. Sato expanded FrameSQL to handle the data from Spanish FrameNet, and he is working on tools to facilitate cross-linguistic comparisons. He attended the Crosslinguistic FrameNet meeting in October.

4.2.4 Publications and Presentations

Lexical Resources Evaluation Conference (LREC), in Lisbon May 28-31, 2004. In addition to a demo and a paper in the main conference on frame-frame relations by Charles Fillmore and Collin Baker, FrameNet staff, in collaboration with researchers from the SALSA project, gave papers at and ran a workshop on "Building Lexical Resources from Semantically Annotated Corpora", attended by roughly 30 people.

Association for Computational Linguistics (ACL), in Barcelona. Collin Baker participated, along with Martha Palmer (PropBank), Manfred Pinkal (SALSA) and Jan Hajič (Czech Dependency TreeBank), in a workshop July 21, 2004, entitled "Beyond Syntax: Predicates, Arguments, Valency Frames and Linguistic Annotation", attended by roughly 40 people.

TreeBanks and Linguistic Theory, in Tübingen, Germany. Collin Baker gave an invited talk "Decorating Trees: A Frame-Semantic Perspective on Adding Semantics to Parse Trees", Dec. 8, 2004.

4.2.5 Applications of FrameNet, S. Narayanan

As a rich, theoretically well founded lexical semantic resource, FrameNet can potentially benefit a variety of Natural Language Processing systems. In 2004, our goal was to make FrameNet (http://www.icsi.berkeley.edu/framenet) usable by making the resource available to researchers in a variety of formats and by demonstrating it's utility for applications in Natural Language Processing. The following specific projects were accomplished to address this goal.

Framenet meets the semantic web OWL (http://www.w3.org/TR/owl-ref/) is a widely used language related to the Semantic Web initiative. The OWL language is being developed as an extension to XML and the Resource Description Framework (RDF). The latest release of the language provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. In 2004, S. Narayanan implemented an OWL ontology for the FrameNet database and with the help of summer intern Matt Gerber built a Java-based translator which creates an OWL version of the FrameNet database. The OWL version and the translator code is now part of all future releases of the FrameNet resource.

FrameNet for Question Answering In 2004, ICSI was part of a team that won a highly competitive AQUAINT Phase II award from the Advanced Research and Development Agency (ARDA) to investigate the use of semantics and inference for Question Answering Systems. ICSI is teaming with the University of Texas, Dallas and with Stanford University on this project. One part of ICSI's role in the project is to build a database of FrameNet Frames for the relevant events and relations of interest to the AQUAINT community. FrameNet is also annotating documents from the AQUAINT database to form a gold standard for members of the AQUAINT community who are trying out machine learning techniques to automate the extraction of semantic relations and roles from textual sources. The FrameNet annotations and database is also being used to build models of inference for Question Answering. The selection of frames and annotation documents involves a continuing cooperation with the Center for Non-Proliferation Studies at Monterey.

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

4.3.1 Episodic Memory for a cognitive system that learns

For several years, Shastri has been developing SMRITI (System for memorizing relational instances from transient impulses), 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 Hippocampal System (HS) as a result of long-term potentiation. The ideas pertaining to episodic memory in SMRITI have been tested in a variety of computational and behavioral experiments.

In 2004, based on his work on SMRITI, Shastri won a new multi-year award under DARPA's Perceptive Assistant that Learns program to develop an episodic memory system for an enduring, personalized cognitive assistant. The goal of the DARPA project is to develop an enduring, personalized cognitive assistant in the form of a software system that can reason, learn from experience, follow instructions, explain its actions, reflect on its experience, and behave robustly in unexpected situations (CALO).

Any cognitive agent, especially an enduring cognitive assistant such as CALO, must have an episodic memory, that is, a memory of past events and situations. An episodic memory records the agent's ongoing experiences and actions and rapidly and retrieves memories of relevant past experiences in appropriate situations. In addition to recording external events, episodic memory also records internal cognitive events. Examples of such internal events are a decision made to perform a specific action at a future time and place (prospective memory), a choice made at a critical intermediate step during planning/problem solving, the detection of conflicts and errors, and the resolution of conflicts and errors.

The specific goals of the effort at ICSI are to:

1. Design and implement an episodic memory system (EM) for CALO,

2. Integrate EM with other components of CALO,

3. Demonstrate that EM has a significant impact on the capabilities and efficiency of CALO.

Specifically, EM will help CALO during task execution and problem solving by reminding CALO - when it is relevant - of past actions that were successful and of actions that it had taken to deal with failures and other contingencies in the past. Thus EM will enable CALO to use its past experience to avoid pitfalls, rapidly adapt and improve its behavior, and make informed decisions and predictions. Moreover, EM will provide CALO with a large set of episodically structured exemplars that can be used to adapt and improve CALO's existing semantic and procedural knowledge by using a variety of symbolic and statistical learning techniques. Furthermore, since EM will enable CALO to replay the states of affair encountered by CALO and the sequence of actions performed by CALO, it will facilitate and support explanation, self-evaluation, and reflection. Finally, EM will serves as a repository of episodic memories shared by the user and CALO. This will enable CALO to understand the user's intent and resolve references to past events made by the user when communicating with CALO.

4.3.2 Evaluating a model of cortico-subcortical loops

In 2004, Narayanan completed preliminary evaluations of his model investigating the role of neuroanatomic loops connecting the frontal cortex to the basal ganglia and thalamus in various aspects of planning and memory. The model was motivated by the following two robust pieces of evidence 1) the pre-frontal cortex plays a key role in various aspects of working memory and executive control and 2) the basal ganglia are closely involved with prefrontal cortex activity. From a functional viewpoint, while damage to the basal ganglia seems to produce cognitive deficits comparable to prefrontal cortex malfunction, teasing out the individual contributions has proven more problematic.

The goal of the evaluation was to flesh out the role of cortical-basal-thalamic loops in planning and executive control. A distinguishing feature of the approach is a fine-grained model of basal-ganglia function that exploits specific component connectivity and dynamics. The model is biologically plausible given current literature on the neurophysiology and disease pathology of the relevant brain regions. A description of the model and preliminary results of applying the model to published behavioral data from Parkinspon's (PD) and Huntington's (HD) subjects on a standard cognitive test (the Wisconsin Card Sorting Task (WCST)) are described in the paper "The role of cortico-basal-thalamic loops in cognition: a computational model and preliminary results", which appeared in the journal NeuroComputing [29].

A significant finding in the 2004 research was that the modeling framework had to be considerably expanded to incorporate neurotransmitter dynamics. EECS Graduate student Joe Makin and Narayanan have been building a stochastic hybrid system model that incorporates both discrete stochastic transitions and continuous state to model the interaction of the four major neurotransmitter pathways impacting the Cortico-basal-thalamic loops. This effort is ongoing and future plans include refining and testing the model in two ways:

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

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

4.4 Analyzing data from the World Color Survey, P. Kay

The World Color Survey (WCS) gathered color naming data from 110 unwritten languages in the 1970s. The data consisted of the names given by each of 25 native speakers to each of 330 perceptually spread color stimuli plus the judgments by the same speakers of the focal (best example) stimuli from this set for each of their color terms. These data had been digitized prior to the current WCS Statistics (WCSS) project but not collected into a single database, which could permit automated cross-language analysis. The conversion to a uniform database was completed by the WCSS project in 2004. In addition several papers analyzing the WCSS data have appeared or are in press. The following are the papers produced by the WCS analysis project at ICSI.

Resolving the question of color naming universals [25, 24] analyze the WCS naming data and demonstrate clear statistical universals in cross-language color naming. [25] shows further that the universals arising from the WCS data on unwritten languages are similar to those arising from the original Berlin and Kay study (1969, Basic Color Terms, Berkeley: U. of Cal. Press) which comprised mostly written languages of technologically advanced societies.

Color naming and sunlight [36] responds to a claim by Lindsey and Brown that excessive ultra-violet radiation, causing premature discoloring of the ocular media (chiefly the lens) accounts for the absence of a lexical green/blue distinction in tropical languages. The disconfirming response shows that the focal points for the 'green-or-blue' terms in tropical languages fall at or close to those for English *green* or *blue* and not where an acquired blue-blindness theory, such as that of Lindsey and Brown, would predict.

Individual and Population Differences in Focal Colors Michael A. Webster and Paul Kay show in [44] that small but significant differences exist across languages with regard to the exact location in color space of the red, yellow, green and blue foci but that these differences are dwarfed by the variation among individuals within languages.

The World Color Survey Database: History and Use [15] details some of the historical and theoretical background of the WCS project and WCSS database and describes its technical nature and potential uses.

Universal Foci and Varying Boundaries in Linguistic Color Categories [37] shows how the boundaries of some color categories as observed in the WCS data, as well as those from one language which has been proposed as a counter to the WCSS universal findings, can be predicted from the universal foci found by WCSS analysis.

References

[1] C. Anderson, P. Domingos and D. Weld. Relational markov models and their application to adaptive web navigation. Proc. KDD-2002. URL:http://www.cs.washington.edu/homes/weld/papers/kdd02.pdf.

[2] S. Atkins, C. Fillmore and C. Johnson. Lexicographic relevance: seeking information from corpus evidence. International Journal of Lexicography, vol. 16, issue 3:251-280, September 2003. Editor T. Fontenelle.

[3] S. Atkins, M. Rundell and H. Sato. The contribution of FrameNet to practical lexicography. International Journal of Lexicography, vol. 16, issue 3:333-357, September 2003. Editor T. Fontenelle.

[4] D. Bailey. When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs. Ph.D. Dissertation, Computer Science Division, University of California Berkeley, 1997.

[5] D. Bailey, J. Feldman, S. Narayanan and G. Lakoff. Modeling Embodied Lexical Development. Computational models of language Understanding, 1997.

[6] C. Baker, C. Fillmore and B. Cronin. The structure of the FrameNet database. International Journal of Lexicography, vol. 16, issue 3:281-296, September 2003. Editor T. Fontenelle.

[7] B. Bergen, J. Feldman and S. Narayan. Embodied verbal semantics: evidence from an image-verb matching task. Proc. Cognitive Science Conference, Boston, August 2003.

[8] B. Bergen, N. Chang and S. Narayan. Simulated Action in an Embodied Construction Grammar. Proceedings of the 26th Annual Meeting of the Cognitive Science Society. Chicago, IL. August 2004.

[9] J. Bryant, J. Feldman and E. Mok. Scaling Understanding Up to Mental Spaces. Second Workshop on Scalable Natural Language Understanding, Boston, May 2004.

[10] J. Bryant. Scalable Construction Based Parsing and Semantic Analysis. Second Workshop on Scalable Natural Language Understanding, Boston, May 2004.

[11] J. Bryant. Recovering Coherent Interpretations Using Semantic Integration of Partial Parses. ROMAND 2004: Robust Methods in Analysis of Natural language Data, Coling 2004, Geneva, August 2004.

[12] N. Chang. A computational model of comprehension-based construction acquisition. Child Language Research Forum, Stanford, CA. May 2004. http://www.icsi.berkeley.edu/~nchang/pubs/CFN-NCPW04.pdf.

[13] N. Chang and O. Gurevich. Context-Driven Construction Learning. Proceedings of the 26th Annual Meeting of the Cognitive Science Society, Chicago, IL, August 2004. http://www.icsi.berkeley.edu/~nchang/pubs/ChangGurevich04.pdf.

[14] B. Chen and P. Fung. Automatic construction of an english-chinese bilingual framenet. In Proceedings of HLT/NAACL, Boston, 2004.

[15] R.S. Cook, P. Kay and T. Regier. The World Color Survey Database: History and Use. To appear in Cohen, Henri and Claire Lefebvre (eds.) Handbook of Categorisation in the Cognitive Sciences, Elsevier, 2005.

[16] J. Feldman and D. Bailey. Layered hybrid connectionist models for cognitive science. NIPS-98, International NIPS Workshop on Hybrid Neural Symbolic Integration, December 4 and 5, 1998, Breckenridge, Colorado, USA, 1998.

[17] J.A. Feldman and L. Shastri. Connectionism. In Encyclopedia of Cognitive Science, Nature Publishing Group, MacMillan, London, 2003.

[18] J. Feldman and S. Narayanan. Embodied Meaning in a Neural Theory of Language. Brain and Language 89 (2004), 385-392, Elsevier Press, 2004.

[19] C.J. Fillmore. Frame semantics and the nature of language. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, volume 280, 20–32, 1976.

[20] C.J. Fillmore, C.F. Baker and H. Sato. Framenet as a "net". In the Proceedings of LREC, volume 4, 1091–1094, Lisbon, ELRA, 2004.

[21] M. Fleischman, N. Kwon and E. Hovy. Maximum entropy models for FrameNet classification. In Proceedings of Empirical Methods in Natural Language Processing, Sapporo, Japan, 2003.

[22] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28.245–288, 2002.

[23] J. Hobbs and S. Narayanan. Spatial Representation and Reasoning. In Encyclopedia of Cognitive Science, Nature Publishing Group, MacMillan, London, 2003.

[24] P. Kay. Color Categories are Not Arbitrary. Cross Cultural Research(2005) 39, 72-8, 2005.

[25] P. Kay and T. Regier. Resolving the question of color naming universals. Proc. Nat. Acad. Sci. 100, 9085-9089 (2003).

[26] T. Murata. Petri nets: Properties, analysis, and applications. In Proc. IEEE-89, volume 77, 541-576, 1989.

[27] S. Narayanan. KARMA: Knowledge-Based Active Representations For Metaphor and Aspect. Ph.D. Dissertation, Computer Science Division, University of California, Berkeley, 1997.

[28] S. Narayanan, C. Fillmore, C. Baker and M. Petruck. FrameNet meets the Semantic Web. First International Semantic Web Conference (ISWC 2003), Sanibel Island, Florida, October, 2003.

[29] S. Narayanan. The role of cortico-basal-thalamic loops in cognition: a computational model and preliminary results. Neurocomputing, 52-54 (2003) 605-614, Elsevier Science, B.V., 2003.

[30] S. Narayanan and S. McIllraith. Analysis and Simulation of Web Services. Computer Networks, 42 (2003), 675-693, Elsevier Science B.V., 2003.

[31] S. Narayanan and S. Harabagiu. Semantic Structures for Question Answering. HLT-NAACL Tutorial, Boston, May 2004.

[32] S. Narayanan and S. Harabagiu. Answering Questions using Advanced Semantics and Probabilistic Inference. Pragmatics of Question Answering Workshop, HLT-NAACL, Boston, May 2004.

[33] S. Narayanan and S. Harabagiu. Question Answering based on Semantic Structures. In Proceedings of the International Conference on Computational Linguistics (COLING 2004), Geneva, August 2004.

[34] S. Narayanan and J. Feldman. CPRM: An expressive probabilistic framework for reasoning about event structure. Submitted for publication, 2004. URL: http://www.icsi.berkeley.edu/∼snarayan/inf-prop.pdf

[35] T. Regier. The Human Semantic Potential. MIT press, Cambridge, MA, 1996.

[36] T. Regier and P. Kay. Color naming and sunlight: Commentary on Lindsey and Brown (2002). Psychological Science, 15, 289-290,2004.

[37] T. Regier, P. Kay and R.S, Cook. Universal Foci and Varying Boundaries in Linguistic Color Categories. To be presented at the 27th Meeting of the Cognitive Science Society, Stresa, Italy, 2005.

[38] L. Shastri. Inference in Connectionist Networks. Cognitive Studies, 10:45-57, 2003.

[39] L. Shastri. The role of temporal coding in the processing of relational information in the mind-brain. In the Proceedings of the 2003 International Joint Conference on Neural Networks, Portland, OR, July 2003.

[40] L. Shastri. Structured connectionist models. In The Handbook of Brain Theory and Neural Networks, II Edition, M. Arbib (Ed), MIT Press, 2003.

[41] L. Shastri. Spreading-activation networks. In Encyclopedia of Cognitive Science, Nature Publishing Group, MacMillan, London, 2003.

[42] L. Shastri and C. Wendelken. Learning structured representations. Neurocomputing, 52-54: 363-370, 2003.

[43] C. Thompson, R. Levy and C. Manning. A generative model for FrameNet semantic role labeling. In Machine Learning: ECML 2003, 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings, ed. by Nada Lavrac, Dragan Gamberger, Ljupco Todorovski, and Hendrik Blockeel, volume 2837 of Lecture Notes in Computer Science, 397408, Springer, 2003. urlhttp://nlp.stanford.edu/ manning/papers/.

[44] M.A. Webster and P. Kay. Individual and Population Differences in Focal Colors. To appear in The Anthropology of Color, ed. by R.E. MacLaury, G.V. Paramei and D. Dedrick. John Benjamins. (projected for 2005 or 2006).

[45] C. Wendelken. Shruti-agent: A structured connectionist architecture for reasoning and decision-making. Ph.D. Thesis, University of California at Berkeley, 2003.

[46] C. Wendelken and L. Shastri. Acquisition of concepts and causal rules in Shruti. In the Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, Boston, MA, July-August 2003.

5 Speech Processing

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

The sections below describe a number of the year's major activities in speech processing. Consistent with our 2003 format we have organized the report in terms of major projects: EARS (large vocabulary speech recognition, metadata extraction, and related tasks); the Meeting Recorder project, whose meeting corpus was released in early 2004 by the Linguistic Data Consortium (LDC); Speaker Recognition; machine query systems (SmartWeb and Clarissa); and continuing joint work with Infineon on exploring physiologically motivated algorithms. We also devote a separate section to a brief update on our completed study of matching speech algorithms to somewhat specialized computer architectures, and to our work on the speech component of a UC Berkeley project seeking to create inexpensive and robust devices to bring information technology to billions worldwide. While this listing includes many of our most significant projects, it is by no means exhaustive, but should provide a useful overview of the major activities in which we have been engaged this year.

5.1 EARS

In 2004 a significant part of our research was focused on the DARPA-sponsored "Effective, Affordable Reusable Speech-to-text" (EARS) program. The goal of this project is to significantly advance the state of the art in multi-lingual speech recognition of both broadcast news and conversational telephone speech. The EARS program consists of two subprojects: Rich Transcription and Novel Approaches. ICSI is a team member on the Rich Transcription project (along with team leader SRI and partner University of Washington) and a lead site for the Novel Approaches project (along with team members SRI, University of Washington, Columbia, and IDIAP in Switzerland; researchers from OGI, a partner in the original project, moved to IDIAP in 2003). Other teams involved in Rich Transcription projects for EARS are Cambridge University, IBM, MIT Lincoln Laboratory, and a BBN-led team that includes LIMSI, University of Pittsburgh, and the University of Washington. LDC is providing data and NIST is under contract to handle the evaluation process. For the Novel Approaches endeavor, the only team outside of the ICSI-led group is a project at Microsoft. The following sections describe some of the major efforts for the year in these two projects.

5.1.1 Rich Transcription

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

ICSI's focus for both broadcast news and conversational telephone speech has been on Core Automatic Speech Recognition (ASR) Algorithms and on Metadata Extraction and Modeling.

In the area of *Core ASR Algorithms*, one major goal was to work to incorporate innovations from our Novel Approaches effort (described below) into the Rich Transcription system. This year we were able to achieve roughly a 10% relative reduction in word error rate on the official NIST evaluation using these methods [46]. In addition to the transfer of Novel Approaches innovations, this year's work in Core ASR also included a number of language modeling explorations. In one of these, we showed that we could reduce the percentage of out-of-vocabulary (OOV) words and the corresponding language model perplexity using information retrieval to select the vocabulary dynamically (without any pre-indexing of relevant documents), and still keeping the vocabulary size the same [4]. Other ASR efforts included work on semantically-motivated language modeling using techniques such as latent semantic analysis, and exploratory work in making better use of the very large amounts of training data available through the EARS program, by training models for "familiar" talkers for broadcast news and by intelligent selection of training subsets for conversational telephone speech.

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

For Diarization, we continued work in improving the performance of a system based on IDIAP's speaker segmentation and clustering algorithm [1]. This year's efforts included modifications of the front-end features, the addition of a speech/nonspeech detector, and refinements to the segmentation routine. The basic system uses a fairly standard agglomerative clustering approach based on a modified version of the Bayesian Information Criterion (BIC), but with the novelty that at each cluster-merging stage it holds the total number of system parameters fixed, thereby obviating the need for the usual BIC penalty term involving a "tweak factor" which generally requires careful tuning for each new application domain. The resulting system is simple, easy to run, and portable. It doesn't require the training of acoustic models on external data and has few tunable thresholds, and so is relatively robust to differences between data sets. For example, when NIST introduced a new development set this year, drawing on a much wider (and more challenging) sampling of news broadcasts, our system achieved an out-of-the-box diarization error rate of 15.7%, very close to its earlier performance of 15.1%, in marked contrast to most other systems where performance degraded dramatically before substantial parameter tuning to achieve comparable results. Of course, the performance of this system does trail that of the best highly-tuned systems that incorporate domain-specific models and parameter settings, and we continue to explore ways of closing this gap using portable, domain-independent techniques. Recent explorations have included alternate stopping criteria for the clustering (using a Viterbi criterion rather than a BIC-based measure) and refined initialization, as well as efficiency improvements such as limiting the list of hypothesized cluster merges. Details of the 2004 diarization system may be found in [44].

For Structural Metadata, the challenge is to detect events that are overtly marked in written text, but "hidden" in the output of a speech recognizer. Currently the primary targets are segmentation of the speech into "sentence-like" units and the detection and repair of various types of speech disfluencies. End goals of this effort include facilitating human readability of recognizer transcripts, as well as aiding downstream natural language processing.

Much of our work in this area [26] has focused on two particular subtasks: the detection of sentence-like units (SUs), and the detection of disfluency interruption points (IPs). The latter are interword locations at which fluent speech becomes disfluent. This includes the interruption point inside an edit disfluency and the starter point of a filler word string. Note that a filler can be either a filled pause (e.g., *uh, um*), a discourse marker (e.g., *you know, like, so)*, or an explicit editing term (e.g., *I mean*).

The following example shows a transcript with metadata marked (using './' for statement SU boundaries, '< >' for fillers, '[ ]' for edit words, and '*' for IPs inside edit disfluencies):

```
and  < uh > < you know > wash your clothes
wherever you are ./ and [ you ] * you really
get used to the outdoors ./
```

Our general approach builds on SRI work on "Hidden Event" modeling [38]. The approach incorporates a language model (LM) capturing sequential information about words and interspersed events, and a prosodic model to predict the probability of an event given duration, pitch, and energy features. In the past year we have made a number of significant improvements and extensions to this approach. In one line of work we improved the prosodic model. Since SU boundary events are much rarer than the nonevents, sampled training sets are generally used to train a decision tree to make it more sensitive to the inherent properties of the events [41]. Liu *et al.* [24] applied bagging and various sampling methods and was able to obtain much more reliable posterior probability estimates from the prosody model. In another set of experiments we extended word-based knowledge, by integrating class-based LMs and LMs trained using auxiliary annotated data with the standard word-based hidden event LM [25, 41]. And experiments in [25] have shown that for the Broadcast News domain, using speaker information obtained from a speaker diarization system is generally better for

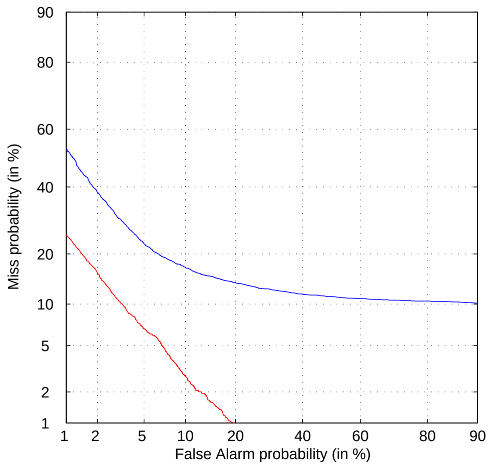


Figure 2: DET curve for SU detection based on confidence predictions for the CTS reference transcript (lower curve) and STT output (upper curve).

SU detection than using speaker change points as derived from the speaker clustering employed for adaptation in speech recognition.

We have also explored alternative modeling approaches, including maximum entropy (Maxent) and conditional random field (CRF) models [25]. Both use features derived from N-grams of words and word classes, binned posterior probabilities from the prosody model, and probabilities from an LM trained using extra text corpora. Combinations of these approaches were used to obtain the SU boundary hypotheses. After SU boundaries were detected, a second step was used to determine the subtype of each SU (statement, question, backchannel, or incomplete) using a Maxent classifier [25]. For disfluencies, three modeling approaches were used. An HMM was used to combine the hidden event LM and a prosody model for IP detection. Heuristic rules were used to find the onset of the reparandum and a separate repetition detector was used to detect repeated words. Second, a Maxent classifier was used to find the IP, and a rule-based approach was used to predict the leftward extent of the edit region. Third, a CRF model was used to detect the edit region and IP jointly. We found that the Maxent and CRF approaches generally outperformed the HMM for the edit word detection task.

System performance degrades significantly using the recognition output rather than reference transcriptions, as indicated by the DET curves in Figure 2.

Detailed analysis [23] has shown that adding textual information, building a more robust prosody model, using conditional modeling approaches, and system combination all yield performance gains. In addition we find that the prosody model, which uses recognition output for purposes of word boundary and phone boundary information, is generally more robust to word recognition errors than are models based on lexical information.

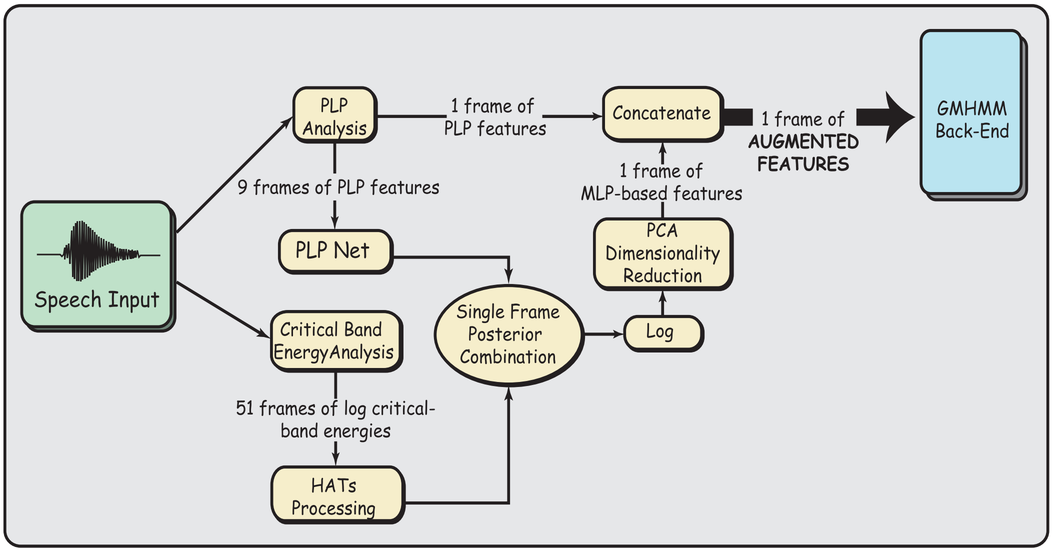

Speech Input
PLP Analysis
1 frame of PLP features
Concatenate
1 frame of AUGMENTED FEATURES
GMHMM Back-End
9 frames of PLP features
PLP Net
1 frame of MLP-based features
PCA Dimensionality Reduction
Single Frame Posterior Combination
Log
Critical Band EnergyAnalysis
51 frames of log critical-band energies
HATs Processing

Figure 3: *Augmenting PLP Front End Features*

5.1.2 Novel Approaches

As noted in previous reports, the objective of this project is to develop a new speech representation to replace the standard cepstral transformation of a local spectral envelope. In 2003 we scaled our methods up from a small vocabulary task to testing on large vocabulary conversational telephone speech [29, 30, 45, 9]. In 2004 we worked with SRI to transfer this technology into a full system for participation in NIST's official Rich Transcription evaluation in the fall, a key milestone for this DARPA-funded project. As noted above, we did this quite successfully, resulting in a 10% relative reduction in word errors for the SRI system [46]. This required considerable effort to handle scaling issues, and in the process we also explored much more deeply alternative neural network architectures.

We have found that posteriors from MLPs focusing on information derived from long time chunks of 500 ms can be effectively combined with posteriors from MLPs focusing on shorter (medium-range) time chunks of 200 ms. The combined posterior goes through further transformation including log, PCA and truncation in the way described in [45], then is concatenated to the traditional features such as MFCC or PLP to form the augmented feature vector, which is passed to a GMM-HMM based speech recognition system. This approach builds on the so-called TANDEM approach first proposed in [15]. The basic feature generation system building blocks are shown in Figure 3.

For both types of MLPs, the output targets are the 46 phones used in SRI Decipher RT-03 system. The MLP focusing on medium term information takes 9 consecutive frames of PLP features, as well as their first and second deltas as inputs. We will henceforth denote this as PLP/MLP. To extract long-term information, we use a variant of the Temporal Patterns (TRAPS) MLP architecture [16, 17] called Hidden Activation TRAPS (HATS) [8, 9]. HATS consists of two stages of MLPs. The first stage extracts phonetically discriminant information from 500 ms of critical band energies, while the second stage merges this information and produces phone posteriors. The phone posteriors from both systems are merged on a per frame basis using a weighted average, where the weights are the inverse entropy of the phone posteriors coming from the corresponding system [28].

From our experience with scaling this approach to use larger amounts of data as well as all the later passes of the SRI recognizer, we were confident that this approach would continue to help when given more data to train with. The challenge, however, was how to train neural nets on an order of magnitude more data (thousands of hours). It was shown in [11] that an optimal ratio of the total number of trainable parameters in an MLP to the total number of training examples is about 1:20. This was 'optimal' in the sense that for a fixed practical limit on the amount of computation time available, there is a ratio that corresponds to the best recognition performance. Since this optimal ratio was roughly constant for different amounts of training data, it is preferable to grow the network size linearly with the increase in training data. Thus the total amount of time for MLP training increases quadratically with the amount of training data. We estimated that it would take more than one year to train our nets on all of the available conversational telephone speech. To speed up training time and yet maintain all the benefits of more data and more parameters, we adopted several modifications to our training recipe.

1. We modified the learning schedule for the nets,
2. rotated the portions of the training data used for each epoch of training, and
3. accelerated the training software by using architecture-specific libraries.

ICSI experiments in the late 1980's showed that early epochs, for which we use a larger learning rate, required less data to get in the vicinity of a good minimum, since the gradient descent error-back propagation algorithm made larger steps in parameter space. However, as we used smaller learning rates for the later epochs, more data helps the algorithm hone in on a good error minimum. We re-introduced this approach, and also adopted a new heuristic, which was to use non-overlapping subsets of the increasing amounts of training data for different epochs. From our experience, having better data coverage gave better results than simply reusing the same data for multiple epochs. By using non-overlapping data in training, the total amount of used training data can cover N/8 + N/4 + N/2 = 7N/8 hours of data, where N = 1200 hours per gender for the CTS training data.

Figure 4 shows the summary of the learning heuristics used to reduce training time.

In addition to modifying the training schedule and employing data rotation, we also took advantage of software upgrades, in particular, using the Basic Linear Algebra Subroutines (BLAS) libraries and the Hyper Threading capabilities of our dual Intel Xeon CPUs. With all of these speed-ups, it took 6 weeks on 4 computers with Xeon 2.8 GHz dual-CPUs to train 4 gender dependent PLP/MLP and HATS nets. Prior to our speedups, extrapolations from smaller experiments suggested a training time of over a year, which would not have been feasible. Finally, feature generation speed was measured as 0.57x realtime on a 3.0 GHz CPU.

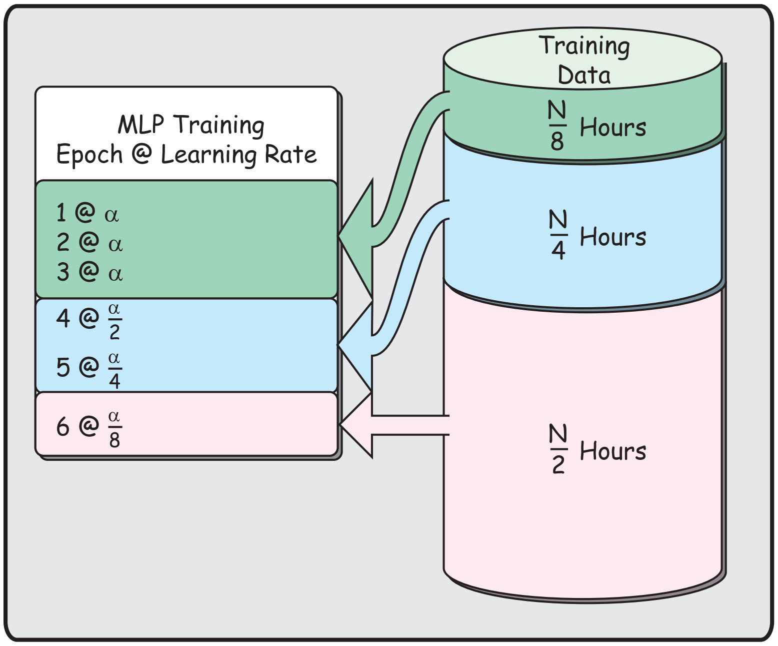


Figure 4: *Streamlined training schedule for large MLP learning task*

Table 1: Word error rate (WER) on RT-04F development and evaluation sets.

	RT-04F Dev			RT-04F Eval		
System	Male	Female	All	Male	Female	All
Baseline	18.1	16.2	17.2	20.2	20.4	20.3
Baseline w/MLP features	16.8	14.2	15.5	19.0	17.7	18.3
Relative change	-7.2%	-12.3%	-9.9%	-5.9%	-13.2%	-9.9%

The baseline for our work is the SRI CTS system as used in the Fall 2003 DARPA Rich Transcription evaluation and later refined for the Fall 2004 evaluation. Two systems were trained: a baseline using standard MFCC (plus voicing) and PLP features, and a contrast system that used MFCCs augmented with Tandem/HATs MLP features. Using the learning heuristics described above, we would incorporate 2100 hours of the data; however, in practice the training sets in the different learning phases were partially overlapped, so that the MLP features were trained on 1800 hours CTS data as described above. This overlap occurred because a portion of the data was more carefully transcribed, so we tended to include it in all of the phases of training. Both systems were tuned on the RT-04F CTS development set (72 conversations) and then tested on the RT-04F evaluation set (also 72 conversations).

Table 1 summarizes all results, split by gender. The overall relative WER reduction on both test sets is identical, 9.9% (2.0% absolute on the evaluation set). However, we also observe that the improvement is almost twice as big for female speakers as for males. This imbalance needs further investigation.

As noted above, we have had to develop a set of heuristics to reduce the rapid growth in training time for the neural networks used in our speech recognition research. However, the database growth is continuing, and the corresponding quadratic growth in training time has not abated; we merely improved the constants involved. Consequently, we also have begun a study of selective sampling to help with this problem. The basic idea is that large speech corpora include many redundancies, as well as outliers and mislabeled samples. More generally, some of the data can be discarded from the training process without any resulting penalty in accuracy. We discarded extremely high entropy speech as computed from posterior estimates from a small MLP, as this indicates low confidence. Low entropy speech is also discarded, with a threshold that is determined experimentally, since it is likely to be redundant. Initial experiments indicate that an MLP trained with 50-60% of the data chosen in this way obtains the same classification accuracy as an MLP trained with all of the data chosen randomly.

In 2004 we developed the tonotopic multi-layered perceptron (TMLP), shown in figure 5, as a competitive alternative to other long-term information capturing systems like TRAPS [18, 17] developed at OGI, or the hidden activation TRAPS (HATS) previously developed at ICSI. As show in the figure, the TMLP incorporates two hidden layers. The first of these is tonotopically organized: that is, for each critical band, there is a disjoint set of hidden units that use the long-term energy trajectory as the input. Thus, each of these subsets of hidden units learns to discriminate single band energy trajectory patterns. The rest of the layers are fully connected to their inputs. Unlike the approach used in HATS and TRAPS, this 4-layer MLP is trained as a single network; in the latter two cases, critical band networks were trained separately with targets such as phones, and then incorporated into a larger system with a merging network. However, like the earlier approaches, it takes advantage of constraints at the input, and incorporates log energies from a very long input sequence (at least .5 seconds). When used to complement traditional short and medium-term front-end features for the recognition of conversational telephone speech, TMLP achieves 8.87% relative WER reduction on Eval2001. This is slightly better than HATS and is a more practical technique in terms of the storage required for training. We also have studied relationships between the dimensionality required for the architecture, particularly in relation to parameters like the number of units at different layers of the network. In particular, we found that the empirically optimum number of critical band hidden units does not grow with increasing training data, but it slightly increases with an increase of parameters. We have also found that the optimal ratio of training frames to parameters is between 20 and 80 and that TMLPs trained in this range have best accuracies when the number of critical band hidden units is between 30-40.

5.2 Speech from Meetings

The automatic processing of speech from meetings – segmentation, transcription, and extraction of information about content and interactions – continues to be a major focus for the Speech group. This year, we made significant advances both in resource creation and in new research. The 75-meeting ICSI Meeting Corpus was at last released by the LDC at the start of the year and now serves as a shared resource for the larger speech

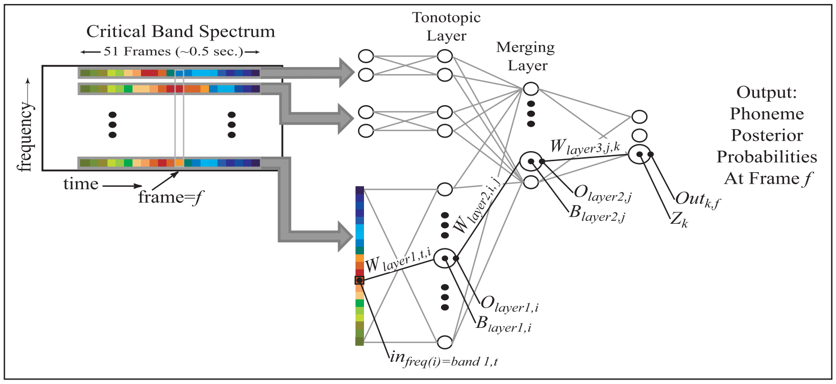


Figure 5: *Tonotopic Multi-Layered Perceptron*

community. We continue to enrich the corpus with additional layers of annotation, this year completing full dialogue act, adjacency pair, and "hot spot" annotations of the meetings, all of which is now being delivered to the LDC for release in 2005. We have also developed a new tool for improved manual transcription and annotation of multi-channel data.

Such data have supported a wealth of research projects this year, including continued development of automatic speech recognition for meetings and participation in NIST's 2004 Meetings Recognition evaluation. With the completion of the hand labeling of dialogue acts, we have also been able to start exploring techniques for their automatic segmentation and classification.

An overview of ICSI's Meeting Corpus, as well as a sampling of the on-going annotations and the research directions it supports, can be found in [20]. All of this work was supported by a combination of sponsors, including: a continuing NSF-ITR project in "Mapping Meetings" ; the Swiss program IM2 (Interactive Multimodal Information Management), funded by the Swiss NSF and managed by IDIAP; and by an award as part of the 15-site consortium AMI (Augmented Multi-party Interaction), an EU Framework 6 program.

5.2.1 Recognition System Building

In 2004 we developed a recognizer for speech from meetings, which was an entry in the NIST Spring 2004 Meeting Recognition Evaluation. This system was developed as a collaborative effort between ICSI, SRI, and UW and was based on SRI's 5xRT Conversational Telephone Speech (CTS) recognizer. The CTS system was modified for the Meetings domain by adapting the CTS acoustic and language models, adding noise reduction and delay-sum array processing for far-field recognition, and adding postprocessing for cross-talk suppression for close-talking microphones. A modified Maximum A Posteriori (MAP) adaptation procedure was developed to make best use of discriminatively trained Maximum Mutual Information Estimator (MMIE) prior models. These meeting-specific changes yielded an overall 9% and 22% relative improvement in word error rate as compared to the original CTS system, and 16% and 29% relative improvement as compared to our 2002 Meeting Evaluation system, for the individual-headset and multiple-distant microphones conditions, respectively. For the evaluation test, the system achieved a word error rate of 34.8% for the personal headset microphones and the multiple tabletop microphone word error rate was 47.0%. We also submitted a contrast system using a more elaborate recognition protocol taking 20 times real-time, rather than under 10. This enhanced system achieved word error rates of 32.7% on the individual headset mics condition, which was the best word error rate reported. On the other required condition, which was the multiple distant mic case, our word error rate of 44.9% was comparable with the leading evaluation system. Details of the Meetings ASR system can be found in [27, 39].

In addition to the evaluation effort, we also were involved in the development of a new Meetings ASR system for the EU Integrated Project AMI (Augmented Multi-party Interaction), a highly-distributed multi-site development effort led by Thomas Hain at Sheffield and using the HTK toolkit as a basis. Our role in this development has been primarily advisory, but also includes contributions to benchmarking, lexicon, and language model development. Finally, we began working to improve the recognizer performance for non-native speakers. The ASR system is trained primarily on Switchboard data, which is all American English speakers, but most test data that we will be using for AMI will be from non-native speakers of the language or from native speakers of non-US dialects (UK, Australian, Indian English). Our approach uses a combination of acoustic adaptation to speech of non-native talkers and pronunciation modification based on linguistically-motivated transformations. We have also explored modifications to the language model to reflect differences in word usage.

In other research that is related to this system development, one of our Finnish visitors, Tuomo Pirinen, completed a project on processing signals from microphone data, using the IDIAP Meeting data to determine Direction of Arrival (DOA) estimation using pairwise time delays between microphones. In particular, he demonstrated a detection method that can minimize the error in this estimation given multiple hardware failures in the distributed microphones [31]. This has potential application beyond speech recognition to other sensor array problems for which location of a sound source is important.

More recently, we have begun exploring the performance of our Diarization system, described under the EARS program above, when ported from the Broadcast News to the Meetings domain. We are looking at ways to make good use of the multiple tabletop microphones, including the use of delay-sum techniques to create a single enhanced far-field channel and the use of the time delays themselves as auxiliary information to assist in speaker segmentation and clustering.

5.2.2 Dialog Act Annotations

As we noted last year, we define a dialog act as the characterization of the function or role of an utterance in the context of the conversation. A set of 58 tags was defined for this work, based on the Switchboard-DAMSL conventions [21] and refined over time by ICSI's annotation team to reflect phenomena observed in the meeting data. The basic utterance types of statement, question, and backchannel (such as "uh-huh") form the primary layer of description, with additional tags providing multiple levels of refinement.

This year we completed a release version of the Meeting Recorder Dialog Act annotations to the ICSI Meeting Corpus, and have made it available to the Linguistic Data Consortium (LDC), who currently distribute the audio and transcriptions for our corpus. The earlier annotations, consisting of over 180,000 hand-annotated dialog act tags from 75 meetings, were augmented to include better adjacency pair information. We added annotation of "hot spots", i.e., regions of particularly high intensity or involvement. We also added explicit speaker labels, fixed minor transcript errors, and a variety of other documented changes. The hot spot annotations were also post-processed and documented for the release. The annotation system and numerous real examples from our data are provided in a detailed manual at http://www.icsi.berkeley.edu/~ees/dadb . The release also includes additional information, such as files specifying the alignment between MRDA segments (dialog act units) and the original corpus punctuation (based on acoustic segmentations).

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

5.2.3 Dialog Act Segmentation and Classification

In 2004, we found that a very simple prosody model improves dialog act segmentation and classification performance beyond that achieved by using word-based information alone [3]. We explored segmentation and classification separately as well, in order to provide a first baseline for these tasks in the ICSI Meeting Corpus. A long-term approach should model these tasks jointly, along the lines of [42], which addressed joint modeling for a much simpler domain. Both problems are difficult and require further work. These problems are even more difficult if one uses far-field audio data, due to added issues of speaker segmentation and tracking. Our initial study also proposed various new metrics for evaluating results. We used simple models, such as a pause-only approach to segmentation, which should show considerable improvement from adding in a larger set of prosodic features (e.g., pitch and energy patterns) in the vicinity of each word, a feasible next step. This would allow comparison to the problem of SU (sentence-like unit) detection in the EARS program [26], since DA segments here have close overlap with the SU segments used in that effort [47]. Note that therefore DA segmentation is generally useful even in the absence of DA classification, since automatic parsing into sentence-like units can aid language modeling, playback of coherent units, and many other tasks.

Classification incorporating both textual and prosodic features in a maximum entropy classifier yielded an accuracy of 81% for 5 DA classes using reference transcriptions, which was reduced to 74% when ASR output was used. Interestingly, results showed considerable similarity to a crudely comparable task in previous research on Switchboard telephone conversations [36].

One difference, however, is that quite unlike previous work on telephone speech, we saw little gain from the modeling of dialog act *context* (previous and following dialog acts) in the ICSI meeting data. Further research is needed to better understand this unexpected result, which could reflect modeling differences, differences in DA definitions, or possibly true corpus differences. The last possibility is the most interesting, since it could suggest fundamental differences between face-to-face multiparty meetings among acquaintances, and two-party audio-only conversations between strangers—in the structuring and constraints on DA sequences. Further work is also needed to learn about the effect of domain on the degradation from errorful word recognition and on the contribution from lexical, prosodic, and context-based knowledge sources. Finally, a long-term goal of our work in this area is to assess the contribution of automatic dialog act modeling to downstream tasks in automatic meeting applications.

5.2.4 Tool Development

The speech group has designed and written a new software tool called "iTranscribe". iTranscribe is used for creating and editing transcriptions of multichannel audio recordings, such as those contained in the ICSI Meeting Corpus. Some of the features of iTranscribe are:

- The ability to read/write MRT-formatted transcript files (MRT denotes "meeting room transcript", the XML format used for the ICSI Meeting Corpus)
- The ability to choose between multiple audio channels for waveform display/playback
- The ability to view and edit transcript text and segment boundaries from multiple audio channels
- The ability to play back audio corresponding to transcript segments

Additionally, since iTranscribe uses the Snack Sound Toolkit (http://www.speech.kth.se/snack), it has all of the features that Snack provides, including the ability to handle multiple audio formats (WAV, AU, AIFF, MP3, NIST/Sphere, etc) and to visualize waveforms, spectrograms, and spectrum sections, and it runs on multiple platforms including Windows, Linux, Macintosh, and Solaris.

iTranscribe is currently in final alpha testing and will begin usability testing in February. We plan to make iTranscribe freely available.

5.3 Spoken Language Systems

An unpiloted Russian cargo craft delivered CLARISSA to the International Space Station (ISS) on Christmas Day, 2004. CLARISSA, developed by a NASA team including ICSI researcher Manny Rayner, is designed to save astronauts time and help them become more efficient by reading station procedures to them. To the best of our knowledge, CLARISSA is both the first spoken dialogue system in space and the first Prolog-based application in space.

Speech recognition and language understanding have been developed using the Open Source REGULUS 2 toolkit [34]. This implements an approach to portable grammar-based language modelling in which all models are derived from a single linguistically motivated unification grammar. Domain-specific CFG language models are produced by first specializing the grammar using an automatic corpus-based method, and then compiling the resulting specialized grammar into CFG form. Translation between language centered and domain centered semantic representations is carried out by ALTERF, another Open Source toolkit, which combines rule-based and corpus-based processing in a transparent way [32]. We also developed a general side-effect free dialogue management architecture suitable for command and control tasks, which extends an "update semantics" framework by including task as well as dialogue information in the information state [33]. Finally, we developed varous compilers, including one to convert XML procedures into the internal form used by the dialogue manager and the one used to organize recording of the output voice.

ICSI is also contributing to another spoken language system project: SmartWeb. This project, led by the German Institute for Artificial Intelligence (DFKI), deals with access to semantic Web services on mobile devices such as mobile phones. Speech input and output are well suited to mobile devices and will be a major focus of SmartWeb. The SmartWeb consortium brings together experts from various research communities: mobile services, intelligent user interfaces, language and speech technology, information extraction, and semantic web technologies. A major part of the project vision is the ability for a user to ask a question using a mobile device and immediately receive an answer based on information drawn from the Web. A demonstration system which can answer questions related to the soccer World Cup is being constructed and should be ready in time for the 2006 World Cup. ICSI staff are involved in the creation of the English-language version of the demonstration system and in the development of speech recognition technology for the overall project. There are two main goals for the speech recognition work. The first is high speech recognition accuracy. The second is sophisticated handling of words that are not in the recognition dictionary. The plan is to pass these out-of-vocabulary to the query engine as strings of phonemes or other subword units, in order to allow query terms to be drawn from an unrestricted vocabulary.

ICSI and our project partners at Sympalog and the Universitaet Erlangen-Nuernberg decided it would be feasible to build both the English- and German-language server-based automatic speech recognition (ASR) for SmartWeb using the Sympalog/Erlangen speech recognizer. The use of the same recognizer platform for both English and German should make it easier to transfer speech recognition technology innovations between project partners during the project. For this reason, David Gelbart of ICSI spent two weeks visiting with project partners Sympalog and Univ. Erlangen-Nuernberg. During and following his visit, the Sympalog/Erlangen speech recognizer was installed on ICSI's computers, and trained on the ATIS English-language corpus. The visit was also an opportunity to work with Christian Hacker and Elmar Noeth of the Univ. Erlangen-Nuernberg on an implementation of the TRAPS/Tandem front end approach for ASR, which ICSI has successfully employed recently [46] to the recognition of conversational telephone speech. Work is continuing and if good results are obtained the technology may be integrated into the SmartWeb demonstrator. Finally, Arlo Faria and David Gelbart of ICSI explored the use of pitch information for vocal tract length normalization (VTLN) for ASR. This work may be used to improve ASR performance in the SmartWeb project.

5.4 Speaker Recognition: Modeling Idiosyncrasies in Speaking Behavior

This project is concerned with the discovery of highly speaker-characteristic behaviors ("speaker performances") for use in speaker recognition and related speech technologies. The intention is to move beyond the usual low-level short-term spectral features which dominate speaker recognition systems today, instead focusing on higher-level sources of speaker information, including idiosyncratic word usage and pronunciation, prosodic patterns, and vocal gestures.

The project goal is two-fold: to conduct fundamental research to discover new speaker-distinctive features and encode them into richer, more informative speaker models; and to evaluate the utility of these feature sets and models for speaker recognition and other speech technology applications. The feature discovery efforts are necessarily exploratory, pursuing both a "knowledge-based" track, building on existing linguistic constructs and guided by insights from psycholinguistics and human performance studies, and a more speculative "data-driven" approach, seeking idiosyncratic "vocal performances" – spectro-temporal patterns with high speaker-characterizing power, independent of linguistic constraints.

Because this is the first substantial effort at ICSI on speaker recognition, we have devoted significant time to establishing basic infrastructure, assembling standard testbeds and data sets, and developing and benchmarking baseline technologies. Subcontractor SRI's support efforts have been instrumental in allowing the ICSI team to ramp up rapidly, supplying much-needed infrastructure and expertise while the ICSI team develops its own. In addition, George Doddington has served as a valued advisor to the project. In fact, an early exercise for the team was the replication of Doddington's fundamental work on "idiolectal" language models [10], demonstrating the substantial speaker-characterizing information contained in speakers' word usage, as modeled via simple n-gram language models.

Moving beyond these start-up exercises in replicating existing technologies, the speaker recognition group is now engaged in a number of new efforts, several of which we describe here.

Text-constrained HMMs for text-independent tasks: It is well-known that text-dependent (password-based or text-prompted) speaker recognition systems achieve far more accurate performance than text-independent systems because they allow more focused models given the word identity. However, many applications of interest – such as speaker indexing of audio archives, background verification during commercial transactions, forensic and security applications – are necessarily text-independent: the speaker's words must be used as found; passwords or prompting are not an option. This project explores the possibility of bringing the more focused modeling of text-dependent systems to bear in the context of unconstrained conversational telephone speech, by concentrating on a small set of words – discourse markers, filled pauses, backchannels – that are likely to arise in conversational speech and likely to be spoken in highly habitual, speaker-characteristic ways. Speaker-independent whole-word hidden Markov models (HMMs) are trained for each of the selected words and then adapted to instances of these words found in the speaker's training data. Test speech is scored by first locating instances of these selected words (via automatic speech recognition), scoring the relevant speech intervals against the trained speaker-independent and speaker-adapted word models, and forming the usual likelihood ratio to obtain a final score. Using only about a dozen words or word phrases ("yeah", "uh-huh", "you know", ...) accounting for only about 10% of the test speech, this system achieved impressive performance on a benchmark test of text-independent speaker recognition, the NIST 2001 Extended Data task. Even though the system still lacks several standard features (such as score normalizations), its performance only slightly trails that of state-of-the-art cepstral-based Gaussian mixture model (GMM) systems, which use all the speech, on this task. Moreover, the keyword HMM system adds new information, combining well with the traditional systems to further drive down error rates, and is highly competitive with an analogous text-constrained system using GMM models of the extracted frames [40] that employs much higher-frequency words. This line of experimentation has been conducted by graduate student Kofi Boakye. His results were presented at the Odyssey'04 conference [6] and formed the basis of ICSI's submission to NIST's 2004 Speaker Recognition Evaluation.

Phone-based speaker recognition systems: Another well-established approach to speaker recognition employs n-gram language models on phone sequences, using the output of a speech recognizer run as an open-loop phone-level decoder [2]. As part of our proposed work, we are interested in the data-driven discovery of speech units better suited than phones to such sequence-based speaker modeling. As a first step in this direction, graduate student Andrew Hatch created a baseline system using the standard phone decoding and bigram modeling used in [2], as well as a related approach modeling phone bigram information via support vector machines (SVMs), in the style of [7]. In both cases, Andy was able to match the published results for these techniques and to dramatically reduce error rates below those reported in the literature by using bigram statistics derived from phone lattices rather than using only the 1-best phone decoding. For example, for the 2003 NIST Extended Data task with target speaker models trained on 8 conversation sides, a bigram SVM using a 1-best phone decoding yielded an Equal Error Rate (EER) of 5.9%, while the phone-lattice SVM yielded an EER of 2.0%. When used in combination with a conventional GMM-based system from SRI that had an EER of 2.6%, the resulting combine system achieved an EER of only 1.4%. (See [13] for further details.) These and other results support the view that lattice decodings provide a much richer sampling of phonetic patterns within speech than 1-best decodings.

Frame sequence explorations: The most speculative work to date involves the search for good matches between speech intervals in training and test data, initially starting with sequences corresponding to word, word bigram, or phoneme tokens as the basis of comparison. The intention is both to capture sequential information about the unfolding of speech over time and to compute a goodness-of-fit between train and test sequences without imposing the usual parametric assumptions by instead employing k^{th}-nearest-neighbor techniques. Work so far has explored several different token types, different match criteria (e.g. fixed frame alignments or dynamic time warping to align tokens), and different scoring metrics (including a novel metric based on "hit counts" for matches exceeding a closeness threshold). This work was inspired in part by the work on sequential non-parametric (SNP) modeling by Dragon Systems in the late 90's [43], but extends it to longer frame sequences – including word or phrase length – given the current availability of larger training and test sets through NIST's Extended Data task. The work was conducted by Research Assistant Daniel Gillick. Graduate student Stephen Stafford, who joined the project this summer, assisted in this effort, developing the dynamic time warping code used in token alignment. While still very much in an exploratory phase, we have already demonstrated the promise of this approach, achieving error rates competitive with a number of more established systems employing higher-level features, such as those using phone n-grams and various prosody-based models. At this time we are still constraining the sequence matching using linguistically-motivated units (phones, words, ...), but we plan to extend these efforts to dynamic sequence expansion anchored at good "match points" and to the sort of variable-length matches employed by unit selection algorithms common to speech synthesis systems. Nonetheless, our current system achieves results that are almost as good as a more standard GMM system, and – given the very different type of information captured – combines extremely effectively with it [12].

Building an open-source GMM system: The dominant model for text-independent speaker recognition systems today is unquestionably the frame-based GMM system [35], which serves as the measure against which new innovations are tested and the basic system with which new information sources are combined. The ICSI team is currently using a traditional GMM system provided by SRI for this purpose. However, it is our intention to phase out this dependency, by creating a competitive GMM system based on open-source components that can then be freely shared with other members of the speaker recognition community. This is in many respects a side effort in terms of the project's technical focus, but one of crucial importance in validating the performance of the novel models emerging from this project and in making the ICSI effort self-sufficient. As the basis for our GMM system development, we are using the Torch toolkit (http://www.torch.ch) developed by IDIAP. Torch provides a highly modular environment for building many types of models, including but moving well beyond simple GMMs (to include HMMs, SVMs, nonparametric models, and others), thus providing good flexibility for future developments. The system construction at ICSI has been primarily an undergraduate project, begun by Shong Yin and now being continued by Shawn Cheng.

5.5 Computer architectures for speech processing

Speech recognition on vector architectures would benefit many applications, including dictation on the desktop, command and control of PDAs and cellphones, and automated call centers using supercomputers. Low power consumption, high absolute performance, and low cost all contribute to the value accrued to vector architectures. To realize these benefits, speech recognition algorithms must be vectorized to run on these platforms.

For Adam Janin's thesis [19], completed in 2004, a vector simulation library was developed to aid in the analysis of speech recognition algorithms on vector architectures. The vector simulation library implements many of the common opcodes found on vector processors, but does not attempt to simulate the fine details of the architectures (cache, chaining behavior, etc.). Instead, the focus of the research is on generating code that vectorizes well on any vector processor.

Of the three major components of ICSI's hybrid speech recognition system, two vectorize quite well. The signal processing component's principal computational bottleneck is the computation of the filterbank, which is typically implemented using a Fast Fourier Transform (FFT). Since the FFT is used in many, many applications for which vector processors are used, most architectures provide some support for FFT computations. The other elements of the signal processing component typically vectorize quite well.

The phone probability estimator used in this work was implemented as a multilayer perceptron (MLP). The computational bottleneck of an MLP is a matrix-matrix multiply. This operation is quite regular, and vectorizes well. However, for optimal performance, fine details of the memory hierarchy must be taken into account. Since such details vary widely, we advocate a generate-and-test approach, where many algorithms are generated automatically, and the fastest is used. Several algorithms were presented that would form the basis for the automatically generated code.

The case of the final component, the decoder, is divided into small and large vocabularies. For large vocabularies, it is desirable to avoid repeatedly computing common prefixes of words (e.g. "four", "fourteen", "forty", "forward"). Also, one can use several methods to avoid altogether the computation of some of the words. For small vocabularies, the savings using these methods are less important, and it is acceptable to simply evaluate each word in full.

Two algorithms were presented for small vocabularies, where every word in the dictionary is evaluated. The first involves batching together words such that the summed length (in states) of a batch is equal to the vector length. The algorithm vectorizes along the state axis of the Viterbi table. The batches are computed using a bin packing algorithm, and all the dictionaries packed quite well. The algorithm itself vectorizes efficiently, and accesses memory minimally. However, it depends on reasonably long vectors, making it unsuitable for some architectures.

The other algorithm batches together words with similar numbers of states. The algorithm vectorizes by word, such that elements of a vector each hold a state from a different word. The algorithm vectorizes well, although it requires extra memory accesses avoided by the algorithm described above. Also, for long vector length architectures, the efficiency can be low, as not all the vectors will be full. For short vector length architectures, however, the efficiency is excellent for all but the smallest dictionary.

For large vocabularies, no method was found that vectorizes efficiently when pruning and tree structured lexicons are used. The base of the problem is that the tree structured lexicons are bushy and unbalanced. The amount of work necessary to arrange for a vectorized operation is nearly the same as just performing the operation directly. Comparisons with the vectorizable small vocabulary systems give an indication of the vector speedup required for a particular dictionary to run more efficiently on vector processor than on a scalar processor. The vectorized small vocabulary system is competitive for all but the largest dictionaries and the highest pruning levels.

5.6 Physiologically motivated speech processing

The human ability to understand speech in challenging situations (such as in the presence of background noise) is superior to the performance of current automatic speech recognition (ASR) systems. This motivates research into ASR techniques which model, or are inspired by, the human auditory system. In collaboration with colleagues at Infineon in Germany, we have performed automatic speech recognition experiments in which a detailed model of the human cochlea and auditory nerve is used for signal analysis in combination with a conventional hidden Markov model statistical back end. To interface the auditory model with the back end we performed simple weighted averages over time and frequency in order to reduce the dimensionality of the auditory model output. This did not outperform conventional mel-frequency cepstral coefficient (MFCC) signal analysis in our experiments, which used the Aurora 2 corpus of speech in noise. This suggests that novel interfacing or acoustic modeling techniques may be necessary to make optimal use of auditory models in ASR. The current auditory model is able to generate realistic output streams of auditory nerve action potentials, and in future work we hope to make use of spiking neural networks to directly process spiking outputs from the model. Also, we hope to extend the auditory modeling in future work to include modeling of the inferior colliculus and cochlear nucleus. To facilitate this planned future work, we created a noise-added version of the ISOLET speech corpus, which we hope will allow us a more rapid experimental turnaround time than the Aurora 2 corpus. This new corpus is currently being fine-tuned and we plan to make it publicly available to the research community once it is finished.

In addition to working with detailed auditory models, we are also interested in applying the principles we observe in the auditory system to conventional techniques in automatic speech recognition. Noting that the mel-warped frequency analysis and logarithmic magnitude employed in the conventional MFCC extraction for ASR can be seen as analogs to the auditory filters and nonlinear compression found in the human auditory periphery, we introduced a simple model of a later stage, namely synaptic adaptation (a mechanism by which signal onsets are strengthened), and integrated it into MFCC calculation. We found that on the Aurora 2 corpus it gave improved performance compared to both plain MFCCs and MFCCs processed with the popular RASTA method [14].

5.7 TIER

Members of ICSI's Speech Group are working to provide speech recognition technology to UC Berkeley's TIER project. TIER (Technology and Infrastructure for Emerging Regions) aims to address the challenges in bringing the Information Technology revolution to the masses in developing regions of the world. Technologies developed for the affluent world and imported to developing regions often fail to address key challenges in cost, deployment, power consumption, and support for semi- and illiterate users. This issue of technology and illiteracy prompted Chuck Wooters and Madelaine Plauche to begin developing a speech recognizer for Tamil, a language spoken by over 50 million people in Southeast India, where illiteracy rates hover around 50% for men and between 60% to 80% for women. Speech recognition, especially in combination with text-to-speech and visual user interfaces, may be key in increasing access to technology to those with limited or no literacy.

During the summer of 2004, Rabin Patra and Sergiu Nedevschi, both graduate students in Computer Science at UC Berkeley, designed a data collection system and collected data on 30 Tamil words (including digits and navigational command words such as “help”,“yes”, and “repeat”) from 8 native speakers at the UCB campus and 22 native speakers in Tamil Nadu, India. Chuck Wooters used this data to build a digit recognizer for Tamil that achieved a 2.7% error rate. Madelaine Plauche examined the data by hand, removing incomplete and mistranscribed data, which further reduced the digit recognizer’s error rate to 1.6%. In addition, she found that the digit recognizer performed equally well whether it was trained on native speakers living in Berkeley or native speakers living in India.

In February of 2005, Madelaine Plauche traveled to three different sites in Tamil Nadu, India to collect more data of native speakers saying digits and command words in Tamil. She hopes to sample the speech of both uneducated and educated speakers, urban and rural speakers, and speakers from three different geographical dialects, to further investigate how dialect and geographic location may affect recognition error rates.

References

[1] J. Ajmera and C. Wooters. A Robust Speaker Clustering Algorithm. Proc. IEEE Speech Recognition and Understanding Workshop, St. Thomas, U.S. Virgin Islands, 2003.

[2] W. Andrews, M. Kohler, J. Campbell, J. Godfrey and J. Hernandez Cordero. Gender-dependent phonetic refraction for speaker recognition. Proc. ICASSP-02, vol. 1, pp. 149-52, 2002.

[3] J. Ang, Y. Liu and E. Shriberg. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. To appear in Proc. ICASSP-2005, Philadelphia, March 2005.

[4] B. Bigi, Y. Huang and R. de Mori. Vocabulary and Language Model Adaptation using Information Retrieval. In Proc. ICSLP’04, Korea, 2004.

[5] J. Bilmes. Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling. In Proc. ICASSP-1998.

[6] K. Boakye and B. Peskin. Text-constrained speaker recognition on a text-independent task. Proc. Odyssey’04 Speaker and Language Recognition Workshop, Toledo, Spain, pp.129-34, 2004.

[7] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones and T.R. Leek. Phonetic speaker recognition with support vector machines. 2003.

[8] B. Chen, S. Chang and S. Sivadas. Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers. In P. Dalsgaard, B. Lindberg, H. Benner and Z. Tan, editors, Proc. EUROSPEECH, Aalborg, Denmark, Sep. 2001.

[9] B. Y. Chen, Q. Zhu and N. Morgan. Learning long-term temporal features in LVCSR using neural networks. In Proc. ICSLP-2004, Jeju, Korea, 2004.

[10] G. Doddington. Speaker recognition based on idiolectal differences between speakers. Proc. Eurospeech'01, vol. 4, pp. 2521-24, 2001.

[11] D. Ellis and N. Morgan. Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. In Proc. ICASSP-1999, 1999.

[12] D. Gillick, S. Stafford and B. Peskin. Speaker Detection Without Models. To appear in Proc. ICASSP-2005, Philadelphia, March 2005.

[13] A. Hatch, B. Peskin and A. Stolcke. Improved Phonetic Speaker Recognition using Lattice Decoding. To appear in Proc. ICASSP-2005, Philadelphia, March 2005.

[14] W. Hemmert, M. Holmberg and D. Gelbart. Auditory-based Automatic Speech Recognition. Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, Jeju, Korea, October 2004.

[15] H. Hermansky, D. Ellis and S. Sharma. Tandem connectionist feature stream extraction for conventional HMM systems. In Proc. ICASSP, vol. III, pp. 1635–1638, Istanbul, June 2000.

[16] H. Hermansky and S. Sharma. Temporal patterns (TRAPS) in ASR of noisy speech. In Proc. ICASSP, Phoenix, AZ, Mar. 1999.

[17] H. Hermansky, S. Sharma and P. Jain. Data-derived nonlinear mapping for feature extraction in HMM. In Proc. ICASSP, Istanbul, June 2000.

[18] H. Hermansky and S. Sharma. TRAPs: Classifiers of TempoRAl Patterns. In Proc. ICSLP-1998.

[19] A. Janin. Speech Recognition on Vector Architectures. UC Berkeley PhD Dissertation, 2004.

[20] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters and B. Wrede. The ICSI Meeting Project: Resources and Research. Proc. NIST 2004 Meeting Recognition Workshop, Montreal, May 2004.

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

[22] D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler and N. Morgan. The Berkeley Restaurant Project, ICSLP-94.

[23] Y. Liu. Structural Event Detection for Rich Transcription of Speech. Ph.D. thesis, Purdue University, 2004.

[24] Y. Liu, E. Shriberg, A. Stolcke and M. Harper. Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection. In Proc. of ICSLP, 2004.

[25] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin and M. Harper. The ICSI/SRI/UW RT-04 structural metadata extraction system. In Proc. of EARS RT-04 Workshop, 2004.

[26] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hilliard, M. Ostendorf, M. Tomalin, P. Woodland and M. Harper. Structural Metadata Research in the EARS Program. In ICASSP 2005, Philadelphia, In Press.

[27] N. Mirghafori, A. Stolcke, C. Wooters, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin and M. Ostendorf. From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System. Proc. ICSLP-2004, Jeju, Korea, October 2004.

[28] H. Misra, H. Bourlard and V. Tyagi. New entropy based combination rules in HMM/ANN multi-stream ASR. In Proc. ICASSP, Hong Kong, Apr. 2003.

[29] N. Morgan, B. Y. Chen, Q. Zhu and A. Stolcke. Scaling up: Learning large-scale recognition methods from small-scale recognition tasks. In Proc. Special Workshop in Maui(SWIM), 2004.

[30] N. Morgan, B. Y. Chen, Q. Zhu and A. Stolcke. TRAPping conversational speech: Extending TRAP/TANDEM approaches to conversational telephone speech recognition. In Proc. ICASSP-2004, 2004.

[31] T. Pirinen and J. Yli-Hietanen. Time delay based failure-robust direction of arrival estimation. Proc. 3rd IEEE Sensor Array and Multichannel signal Processing Workshop, Barcelona, Spain, July 2004.

[32] M. Rayner, B.A. Hockey, J. Hieronymus, J. Dowding and G. Aist. An Intelligent Procedure Assistant Built Using REGULUS 2 and ALTERF. Proc. 42nd Annual Meeting ACL, Sapporo, Japan, 2003.

[33] M. Rayner and B.A. Hockey. Side Effect Free Dialogue Management in a Voice Enabled Procedure Browser. Proc. ICSLP'04, Jeju Island, South Korea, 2004.

[34] M. Rayner, B.A. Hockey and P. Bouillon. Building Linguistically Motivated Speech Recognisers with Regulus. Proc. 42nd Annual Meeting ACL, Barcelona, Spain, 2004.

[35] D.A. Reynolds, T.F. Quatieri and R.B. Dunn. Speaker verification using adapted gaussian mixture models. Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[36] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer and C. Van Ess-Dykema. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? Language and Speech, 41(3-4), 439-487, 1998.

[37] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang and H. Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. Proc. SIGDIAL-2004, April-May 2004, Boston.

[38] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tür. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. Speech Communication, eds. T. Robinson and S. Renals, vol. 32, 1-2, 127-154, Sep, 2000.

[39] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin and M. Ostendorf. Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System. Proc. NIST 2004 Meeting Recognition Workshop, Montreal, May 2004.

[40] D.E. Sturim, D.A. Reynolds, R.B. Dunn and T.F. Quatieri. Speaker verification using text-constrained gaussian mixture models. Proc. ICASSP-02, vol. 1 pp. 677-80, 2002.

[41] M. Tomalin and P. Woodland. Advances in structural metadata for Eval04 at CUED. In Proc. of EARS RT-04 Workshop, 2004.

[42] V. Warnke et al. Integrated Dialog Act Segmentation and Classification Using Prosodic Features and Language Models. Proc. ICASSP-01, 2001.

[43] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel and L. Gillick. Speaker recognition on single- and multi-speaker data. Digital Signal Processing, vol. 10, no. 1-3, pp. 75-92, 2000.

[44] C. Wooters, J. Fung, B. Peskin and X. Anguera. Towards Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System. Proc. EARS RT-04F Workshop, Palisades, New York, November 2004.

[45] Q. Zhu, B. Y. Chen and N. Morgan. On using MLP features in LVCSR. In Proc. ICSLP-2004, Jeju, Korea, 2004.

[46] Q. Zhu, A. Stolcke, B. Y. Chen and N. Morgan. Incorporating Tandem/HATs MLP Features into SRI's Conversational Speech Recognition System. Proceedings of the EARS RT-04F Workshop, Palisades, New York, November 2004.

[47] NIST Website. RT-03 Fall Rich Transcription. http://www.nist.gov/speech/tests/rt/rt2003/fall/.

6 Other Activities

6.1 Information Society Program (formerly the Berkeley Center for the Information Society)

The Berkeley Center for the Information Society was a research center started at the International Computer Science Institute to bring computer science and social sciences together. Although this interdisciplinary connection is still seen as very critical for the next stage of the technological and social development, the continuing ICSI effort is no longer considered a separate group.

The Center's initial focus areas included: (1) challenges of the global information society and different models of responding to it; (2) the use of the open-source model for social projects; and (3) enhancing equal opportunities within IT. All of these efforts continue, but the comparative studies of item (1) are no longer centered at ICSI.

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

BFOIT continues industrial relations with events like the "New Technologies Day" at Microsoft This grew from 55 to 68 participants and one of the vice-presidents asked us to keep expanding the program. Several parents chaperoned the all-day event and they and the students were very pleased with the event which involves students in North Carolina, Texas and Silicon Valley. There are talks emerging around starting another BFOIT unit in the San Jose area, initiated by an active parent in the current program. BFOIT is looking to expand and export its successful programs. Guy Haas has been helping Albany High School adapt our course materials and there are also discussions with Emery High School. An NSF proposal for extensive outreach is under review.

In 2003, BCIS undertook a study of Civil Communities of Practice for Sitra, the Finnish development agency. The resulting report is available and is forming the basis for specific efforts in Finland, the Central Valley and elsewhere. In 2004, this effort turned to implementation and deployment of a pilot system. The Communities of Practice Environment (CoPE) is a novel software platform for supporting cooperative effort among formal and informal groups of people who may be separated in time, space, and language. Essentially all previous group-oriented software has focused on one of two approaches: 1) systems for institutional settings, probably because institutions have the most resources; and 2) groups of small scope and limited interaction, for which extremely process-rigid free products (such as Google or Yahoo Groups) are provided. But there are many thousands of voluntary organizations that can benefit from software that helps them work together and achieve concrete goals beyond what is available through existing products. We have developed such a system, called CoPE, implemented it, and deployed it in a variety of user environments. CoPE does not depend on technically sophisticated users or on supporting institutional structures in fact, CoPE itself can provide the mechanism for groups to establish their preferred working style and institutional memory. The two pilot deployments involved very different user groups. The first group (CVP) consists of people who are widely dispersed geographically, are not experienced computer users, and are not all comfortable in English. The organization is a public one and there is contention for resources and other controversy within the group. This group has been using the CoPE since Fall 2004 and has exploited the archiving capability to organize past as well as ongoing activities.The second pilot group (CogFac) comprises faculty members at a major research university. They are all sophisticated users and some are computer scientists who do research on software.

Early in 2004, Prof. Anna Saxenian of BCIS was appointed Dean of the UCB School of Information Management and Systems. Also, Prof. Steve Weber was appointed as the Director of the UCB Institute of International Studies. David Thau, a SIMS doctoral student working on the CoPE project has also been accepted to the UCB Boalt Law School. This, plus the continuing cooperation with Citris and the ICT4B project ensures strong campus participation in BCIS activities.

6.2 Robust Video Compression based on Distributed Source Coding techniques

In this project, headed by EECS Prof. K. Ramchandran, and staffed by Jiajun Wang and Abhik Majumdar, we study the problem of enhancing a baseline MPEG system through the use of a distributed source coded auxiliary channel. Predictive video coding suffers significant quality degradation in the event of channel loss due to prediction mismatch between the encoder and decoder, also called "drift". As shown in Figure 6, in our system, extra information is coded using a Wyner-Ziv framework and sent over a low-rate auxiliary channel as a second description of the predictively coded video. The erroneous predictive decoder reconstruction then serves as side-information for the auxiliary channel decoder to obtain a higher-quality reconstruction. Correlation between the erroneously reconstructed frame and the original frame is estimated using a modified version of the concepts detailed in [1]. Assuming an independent packet erasure model, we develop a quantitative analysis to optimize drift reduction using the setup of Figure 6. Further our design features backward compatibility such that if the decoding client only has an MPEG/H.26x decoder, it can still decode the MPEG/H.26x bitstream after throwing away the auxiliary channel bitstream.

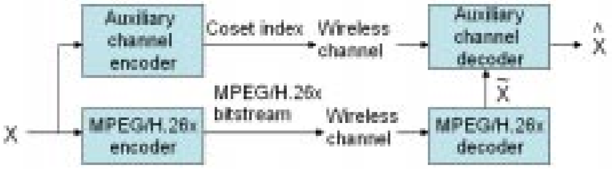
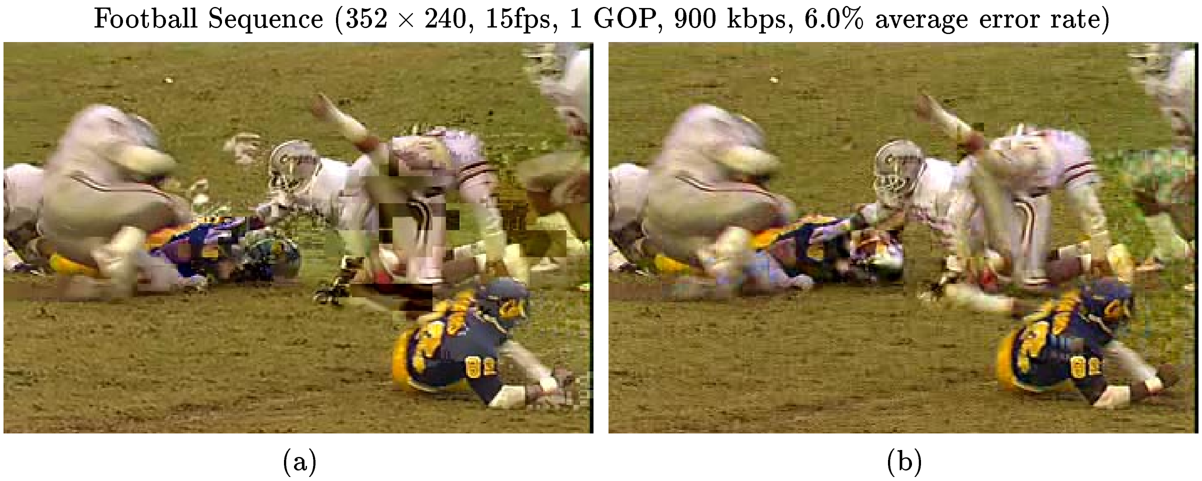


Figure 6: System block diagram. The conventional predictive decoder reconstructs the input $\mathbf{X}$ as $\tilde{\mathbf{X}}$. $\tilde{\mathbf{X}}$ serves as side-information to the auxiliary channel decoder which outputs $\hat{\mathbf{X}}$ as the final reconstruction.

This project was started last year when we developed the basic algorithm. This year's highlights include:

- Completed the implementation of a working system.
- Improved the performance of the auxiliary channel in unicast setting by developing a rate-control algorithm.
- Optimally allocated rate between distributed coded data and Forward Error Correction codes (FEC) to take advantage of FEC's error prevention capacity.
- Expanded the use of auxiliary channel to a multicast setting by adopting broadcast source coding concepts from multi-user information theory in [2, 3]. Homogeneous receivers are optimally satisfied in a rate-distortion efficient manner.

Extensive simulations were carried out on a CDMA 2000 1x wireless network using simulators obtained from Qualcomm. The auxiliary channel implementation proved to outperform the scheme where only FEC is used by 2-4 dB (See Figure 7) and outperform the scheme where MPEG baseline is used by 4-7 dB. We are currently in the process of technology transfer to Qualcomm Inc. and pushing the proposed algorithm into



Football Sequence (352 × 240, 15fps, 1 GOP, 900 kbps, 6.0% average error rate)

(a)

(b)

Figure 7: (a) Frame 10 of the football sequence when only FEC is used on H.264 stream and (b) when distributed source-coded data plus FEC on H.263+ stream is used to prevent/correct drift respectively.

the 3GPP wireless network standards for mobile video broadcast. Below is a list of publications that came out of this project:

- A. Majumdar, J. Wang, K. Ramchandran and H. Garudadri, "Drift Reduction in Predictive Video Transmission using a Distributed Source Coded Side-Channel," in *Proc. ACM Multimedia*, 2004.

- J. Wang, A. Majumdar, K. Ramchandran and H. Garudadri, "Robust Video Transmission over a Lossy Network Using a Distributed Source Coded Auxiliary Channel," in *Proc. Picture Coding Symposium (PCS)*, 2004.

- J. Wang, A. Majumdar and K. Ramchandran, "On Enhancing MPEG Video Broadcast over Wireless Networks with an Auxiliary Broadcast Channel," to appear in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 2005.

References

[1] R. Zhang, S. L. Regunathan, and K. Rose. Optimal intra/inter mode switching for robust video communication over the Internet. In Proc. 33rd Ann. Asilomar Conf. on Sig. Syst. Comp., 1999.

[2] C. Heegard and T. Berger. Rate distortion when side information may be absent. IEEE Trans. Inf. Theory, vol. 31, pp. 727–734, Nov 1985.

[3] Y. Steinberg and N. Merhav. On Successive Refinement of the Wyner-Ziv Problem. Submitted to IEEE Trans. Inf. Theory.