**PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE:**
- NSF 00-126
- 04/10/01

**FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S):** (Indicate the most specific unit known, i.e. program, division, etc.)

**TITLE OF PROPOSED PROJECT:**

**REQUESTED AMOUNT:**

**PROPOSED DURATION (1-60 MONTHS):**

**REQUESTED STARTING DATE:**

**SHOW RELATED PREPROPOSAL NO., IF APPLICABLE:**

**CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW:**

**PI/PD DEPARTMENT:**

**PI/PD POSTAL ADDRESS:**

**PI/PD FAX NUMBER:**

**NAMES (TYPED):**

**PI/PD NAME:**

**High Degree**

**Yr of Degree**

**Telephone Number**

**Electronic Mail Address**

---

**International Computer Science Institute**

1947 Center Street

Berkeley, CA 94704 1105

**ITR/PE+SY: Mapping Meetings: Language Technology to make Sense of Human Interaction**

**$3,911,186**

**48 months**

**09/01/01**

**0105983**

**BEGINNING INVESTigator (GPG I.A)**

**DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C)**

**PROPRIETARY & PRIVILEGED INFORMATION (GPG I.B, II.C.6)**

**NATIONAL ENVIRONMENTAL POLICY ACT (GPG II.C.9)**

**HISTORIC PLACES (GPG II.C.9)**

**SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.C.11)**

**VERTEBRATE ANIMALS (GPG II.C.11) IACUC App. Date**

**HUMAN SUBJECTS (GPG II.C.11) Exemption Subsection or IRB App. Date**

**INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED**

**HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.E.1)**

---

<table>
<thead>
<tr>
<th>Name</th>
<th>Degree</th>
<th>Year</th>
<th>Phone Number</th>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelson Morgan</td>
<td>Ph.D.</td>
<td>1980</td>
<td>510-666-2931</td>
<td><a href="mailto:morgan@icsi.berkeley.edu">morgan@icsi.berkeley.edu</a></td>
</tr>
<tr>
<td>Daniel P Ellis</td>
<td>Ph.D.</td>
<td>1996</td>
<td>212-854-8928</td>
<td><a href="mailto:dpwe@ee.columbia.edu">dpwe@ee.columbia.edu</a></td>
</tr>
<tr>
<td>Katrin Kirchhoff</td>
<td>Ph.D.</td>
<td>1999</td>
<td>206-221-5476</td>
<td><a href="mailto:katrin@ee.washington.edu">katrin@ee.washington.edu</a></td>
</tr>
<tr>
<td>Mari Ostendorf</td>
<td>Ph.D.</td>
<td>1985</td>
<td>206-221-5748</td>
<td><a href="mailto:mo@ee.washington.edu">mo@ee.washington.edu</a></td>
</tr>
<tr>
<td>Andreas Stolcke</td>
<td>PhD</td>
<td>1994</td>
<td>510-666-2969</td>
<td><a href="mailto:stolcke@speech.sri.com">stolcke@speech.sri.com</a></td>
</tr>
</tbody>
</table>

---

**Received:** 04/10/2001

---

**EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN):**

943024996

**NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE:**

International Computer Science Institute

**ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE:**

International Computer Science Institute

1947 Center Street

Berkeley, CA. 947041105

**NAME OF Awardee ORGANIZATION CODE (IF KNOWN):**

400093900

**ARWDEE ORGANIZATION CODE (IF KNOWN):**

**IS Awardee ORGANIZATION (Check All That Apply):**

- FOR-PROFIT ORGANIZATION
- SMALL BUSINESS
- MINORITY BUSINESS
- WOMAN-OWNED BUSINESS

---

**IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY?**

YES

NO

**FILE LOCATION:**

187909478
Mapping Meetings:  
Language Technology to Make Sense of Human Interaction  
ITR/PE+SY

Meetings—situations in which a number of people have a face-to-face discussion—are an essential part of every enterprise or organization, and may even be said to be one of the quintessential human activities, combining verbal communication with group social dynamics. It is conspicuous, however, that in contrast to many other communicative activities, meetings have until now experienced little impact from the development of information technology. Given the importance and ubiquity of meetings, this is bound to change. In particular, recent rapid advances in natural language processing, information extraction and speech recognition have set the scene for a new genre of meeting support and analysis technologies.

This project aims to develop systems for finding information in recorded meetings. The central metaphor is a “meeting map”, an abstract concept encompassing both a structured representation of a particular description of a meeting, and interactive user interfaces for accessing and exploring that information. The concept of a map is particularly apposite because, like traditional maps, meeting maps may display one or several quite different kinds of data about a meeting, they can exist or be rendered at a variety of levels of detail or scales, and their primary purpose is to aid in navigation amongst the represented material.

The range of information that could be usefully included and presented in meeting maps is practically unlimited. The proposed project will have an initial focus on two, contrastive, broad categories of maps:

- **Content maps** portray the subject matter of meetings, the topics discussed, and any decisions made or other substantive goals achieved in the meeting.

- **Interaction maps** are concerned with the social dynamics of the meetings, identifying the roles and relationships of the participants, and the ‘shape’ of the meeting in terms of levels of involvement and concurrence.

Development of meeting maps will require new technologies in several areas. At the lowest level is the automatic acoustic processing needed to extract categorical descriptions from the multichannel recordings of real meetings that are being collected for this project. Research into acoustic processing will include the development of speech recognition engines best suited to information extraction tasks, automatic prosody extraction to find phrase and topic boundaries, and approaches to automatic speaker turn detection and clustering that are robust to the variety of signals observed in meeting recordings. Given these first-level analyses of the raw data, content maps will be built through information-extraction techniques based on topic tracking, key-phrase detection, and automatic measures of content salience. Interaction maps will require novel automatic analysis and classification of dialog acts, speaker roles, disagreement/consensus and decision events as they occur in meetings. In constructing the recognition algorithms and determining the elements of the maps, this project will build on analyses of at least 100 hours of transcribed data currently being collected at ICSI and UW.

These two types of maps will be used for generating different types of textual summaries. The very different styles of literal speech transcripts and conventional summary prose make summarization of meeting data a particularly significant new challenge, requiring extensive investigation of new analysis and generation techniques. In addition, the project will investigate techniques for generating maps (and the resulting summaries) dynamically using query-dependent language models and coupling between the information extraction and recognition processes.

The project will be evaluated through a variety of standard objective measures for components such as speech recognition and information extraction, expert evaluations of automatic summaries and other outcomes, and/or task evaluations. Meeting map data will be presented to users via off-the-shelf visualization products; the focus of the project is on the underlying data representation, not the details and mechanics of the information visualization.

As is required for an undertaking of this scope, the project brings together 12 senior researchers from engineering and social science disciplines at 4 research institutions. Because of the range and ubiquity of meetings—a term that can span chance encounters at the water-cooler through to highly formalized legal proceedings—it is very timely to be advancing the possibilities of automatic meeting analysis. Meeting maps respect the richness and diversity of information and interpretations present in meeting scenarios, and should form the foundation of entirely new technology of human-to-human interaction support.
1 Introduction

Meetings play a crucial role in our social, economic, and political lives. Despite email, cell phones, and other forms of technology-mediated communication, there is still no serious substitute for the good old-fashioned notion of bringing people together for face-to-face discussion. For many purposes, meetings also have the status of documents, in that we often wish to preserve them for posterity, for later reference, and for cross-reference from other meetings and (e.g., written) documents. Keeping records of meetings is desirable for a multitude of reasons—from legal, social and educational, to commercial. Simply recording a meeting, however, is of remarkably little use for most real-world purposes. For example, a person who misses a meeting may want a summary of key points, or a meeting participant may want to search a meeting to recall some specific discussion of a subtopic. A third person may be interested in whether or not there was argument during a particular meeting. But all of these users share a common constraint: none of them wants to wade through a recording or transcript of the entire meeting.

The goal of this project is to automatically process real meeting recordings in ways that will allow people to access them for such purposes. Existing technology provides a starting point in achieving this goal. Automatic speech recognition (ASR) can be used to generate approximate transcripts of the words that are spoken; information retrieval (IR) techniques can aid in searching for relevant content; and recent advances in question answering might help in finding specific information. However, current technology is severely limited when it comes to processing real meetings. ASR is fraught with high error rates for conversational speech (30-50%), and most IR technology is text-based: it presumes accurate transcripts, and that all relevant information can be captured in the written word. Furthermore, there has been very little work on the automatic analysis of the highly informal style of speech that typically occurs in meetings. An excerpt of a meeting about speech transcription standards and tools is given below, and illustrates common phenomena such as disfluencies (-), overlaps (/), interruptions, and misunderstandings.

A: Ok. So that means that for each utterance, .. we'll need the time marks.
E: Right. A: the start and end of each utterance.
[afew turns omitted]
E: So we - maybe we should look at the um .. the tools that Mississippi State has.
D: Yeah.
E: Because, I - I - I know that they published .. um .. annotation tools.
A: Well, X-waves have some as well, .. but they're pretty low level .. They're designed for uh - D: phoneme A: for phoneme-level D: transcriptions. Yeah.
J: I should -
A: Although, they also have a nice tool for - .. that could be used for speaker change marking.
D: There's a - there are - there's a whole bunch of tools
J: Yes. / D: web page, where they have a listing.
D: like 10 of them or something.
J: Are you speaking about Mississippi State per se? or
D: No no no, there's some .. I mean, there just - there are - there are a lot of / J: Yeah.
J: Actually, I wanted to mention - / D: (?)
J: There are two projects, which are .. international .. huge projects focused on this kind of thing, actually .. one of them's MATE, one of them's EAGLES .. and um. / D: Oh, EAGLES.
D: (?) / J: And both of them have
J: You know, I shou-, I know you know about the big book.
E: Yeah.
J: I think you got it as a prize or something.
E: Yeah. / D: Mhm.
J: Got a surprise. {laugh} {J. thought ”as a prize” sounded like ”surprise”}

A main goal of this proposal is to transform collections of meeting information into a form where the most important information can be rapidly, clearly, and accurately conveyed to a user. As seen from the above transcription, it is somewhat difficult and wearisome to read through textually represented meeting material. Moreover, the above transcription was over a duration of only 72 seconds. Searching through many hours of transcribed meeting material in this way, even when there are no word transcription errors (as above) and even when prompted beforehand about the meeting’s topic, would be extremely time consuming.

Meeting material, by definition, consists of multiple people speaking with each other about a single (or a small number) of topics. Some meetings are quite structured, and there are formal systems to ensure that a meeting follows

---

1For this project we have chosen to focus on audio data, because despite the interesting aspects that video can add, there are large challenges already facing us for the audio processing. Furthermore, there are many important meeting contexts for where visual records are simply not available (e.g., teleconferences, radio panel shows, impromptu meetings). The converse—meetings that exist purely as visual events with no soundtrack—are much harder to imagine.
rules of orderly conduct. For example, Robert’s rules of order [39], first established in 1876, are specific laws where motions are made, seconded, assignments are given to the floor, participating members vote, and so on. Other meetings are quite unstructured. They consist of multiple people interacting in a relaxed conversational speaking style that includes word repetitions, disfluencies, and laughter. Meetings often have multiple people speaking simultaneously and (amicably or not) cutting each other off. In such cases, there exists a rapidly changing meeting dynamic based on the feedback participants receive from other members of the group. This can have a significant effect on the words that are used and the style and timing of disfluencies. In either case, because such social dynamics exist in meetings, it is much more difficult to determine the main meeting thrust by reading through textually transcribed meeting material than it would be by reading through normal written text.

Furthermore, meetings are inherently parallel. Unlike the auditory domain, where we are accustomed to handling multiple simultaneous speakers, a textual representation only poorly represents concurrently spoken streams. Text is inherently serial and there is no real provision for conveying cases where multiple people are simultaneously interacting with each other. Text, moreover, is not set up to capture rapid feedback amongst multiple speakers (especially in the case of unstructured meetings).

Ideally, meeting information would be summarized to allow users to quickly access the information of interest to them. Such a facility would support inquiries such as, “what was the main point of person X”, or “did person X and person Y agree”, or “did this meeting pertain to subject Z, and who said the most about it”, and so on. Clearly, a raw presentation of the transcribed data as given above would be an extremely tedious way to answer such queries. As a result, standard question answering, in which the user is pointed to a passage containing the answer—itself a difficult problem – would be of uncertain value for meeting transcripts.

What is needed is a tool that presents the underlying information of a meeting in a structured and coherent way so users can rapidly determine if a meeting matches their interests and efficiently extract the particular items of interest to them. The research we propose has three main thrusts. First, using the metaphor of a map, as described in the next section, we will develop techniques to extract and represent meeting content and social interactions among the participants. Second, we will develop methods for summarizing this information to users in multiple, flexible formats and at different levels of granularity. Finally, we will investigate dynamic extensions of these techniques to allow rapid and query-dependent access to the meeting data.

2 Meeting Maps

A meeting map is a structured representation of a related set of properties of a given meeting that can be used to navigate through the available information (e.g., with the goal of finding a specific bit of information). Just as geographical maps come in different varieties (topographical, political, ethnic, weather), different meeting maps can be used to represent properties such as content, emotional state, and social dynamics. Although the possible types of meeting maps form an open list, in this effort we will focus on two types of maps: content maps (e.g. topics, information status) and social interaction maps (e.g. consensus/disagreement). Like geographical maps, our maps will also vary in resolution. For example, a content map at the coarsest level might contain a succession of topics, identified by keywords; at a finer level however it could contain much more detail.

Meeting maps may be combined in numerous ways to convey and summarize information for human inspection and exploratory data analysis. Different maps of the same meeting can be superimposed to identify patterns that involve different aspects of a meeting. For example, by combining a content map showing salient utterances and a social interaction map identifying the speakers, one can determine if the contributions in a meeting were balanced or dominated by a subset of speakers. Furthermore, maps of different meetings can be cross-referenced, e.g., via links to recurring topics. The mapped information can be accessed either graphically or through a formal query language.

Meeting maps need not be static and are not necessarily precomputed independently of the application in which they will be used. Different users will need to obtain different information from a meeting representation; furthermore, information needs may change while navigating an existing map. It may sometimes be impossible to predict precisely what kind of information a user will be looking for. A particular emphasis of our research will be on the processing techniques for the construction of user-oriented, task-focused and dynamic maps.

As an example, consider the different maps that might be generated for an hour long meeting of the ICSI speech
The meeting began with a discussion of the forms that will be used to get information about the participants in a meeting. Discussion centered around the form of the questions and possible answers that should appear. Most time was spent on one form before moving to the questions that should appear on a simpler form. The meeting closed with a technical discussion of errors in the speech recognition software.

This is an indicative summary, providing a description of topics, and would be useful for an end user to determine whether the particular meeting is of interest and deserving of more detailed examination.

Another high-level view of the meeting might label segments by topic in outline form. The map would provide a time alignment of the key words to the actual speech signal, organized hierarchically to allow expansion or compression to different levels of detail as shown in Figure 2. The figure shows, for example, that the first subsegment within the “Form” topic focuses on a discussion of one question that is proposed to appear on the form and the possible answers that the form should include. Information associated with a particular segment of a meeting might include a set of terms that occurred most frequently within that topic segment, the meeting participants and their roles, as well as the locations during the meeting where disagreements occurred and an indication of whether any consensus was reached. These elements would also be linked to time points in the speech signal, making it possible for a user to play back particular segments of interest. (Note that, even with perfect summarization or transcription, one might want to listen to portions of a meeting to gauge the emotional quality of speakers’ contributions, to be able to later identify their voices, etc.)

In the particular segment illustrated, there is disagreement among the participants as to the choices that should be given for the answer. A meeting map of this segment of the meeting might provide an informative summary, such as

In this segment, participants discuss one question on educational background and the possible choice of answers. They disagree about the answer choices that should be provided with some (e.g., A) arguing for more fine-grained resolution in the answer and others (C, L, N) arguing for less (e.g., student/non-student). Finally, they agree on the categories undergrad, graduate student, post-doc, professor, and a last choice with the words “visiting post-Ph.D. researcher”.

which gives the end user information about the viewpoints taken and the final decision reached. Such a meeting map could be useful in allowing the end user to understand the outcome of the meeting, without having to read the entire transcript.

Producing an informative meeting map such as this is very hard. The excerpt in Figure 1 shows the words and partial phrases that the participants use to convey two viewpoints. The system must be able to identify that consensus is reached when L says “I think that’s fine.” and after discussion of the sticking point (how to distinguish between professors and others who have a Ph.D.), “then I think what you have is great.” Here, the map might assign extracted terms to viewpoint, showing those terms that are used by the participant (A) who advocates more fine-grained choices in the answer versus those used by the participants (C, J, L) who seem to question the need for fine-grained distinctions.
(see Figure 2). To generate the full summary above, the system must use the extracted features (terms, topic segments, identified viewpoints) and phrases from the transcript itself (e.g., “... were you suggesting that it be more fine-grained than this...”, “Five things that you can circle undergrad, grad, post-doc, professor, and then other ...”).

A significant portion of the project will be concerned with producing a summary along with the characteristics of the meetings (topic segments, topic keywords, frequent terms, participants, points of disagreement) that can be provided to a visualization component for display, either in terms of time-based representations of the meeting or on other dimensions. We will demonstrate how the extracted information can be used to create a meeting map through display in an off-the-shelf visualization tool, since details of the graphical visualization are not a primary focus of the project.

3 Related Work

Projects related to transcribing and analyzing meetings are in progress elsewhere [55, 56, 22]. Our proposed research is unique in its focus on producing multiple views in meeting maps of the same dialog. The ability to produce a coherent summary of the dialog, in particular, is a research task that has received very little attention to date. Most work on open domain summarization of speech has focused on analysis of the input as opposed to generation of the summary; identification of dialog features and important content terms are key contributions of research to date. For example, research on summarization of broadcast news in Informedia [54] has developed technologies to segment news into topically related stories, extracting list of key terms from each topic which serve as a sort of summary. Similarly, work at AT&T Research on summarization of broadcast news and documentaries provides indexing of speech for browsing tasks [18, 58] and identification of speaker types to augment a speech transcript with structural information as a first step towards summarization [3]. Others [52] stipulate that only language which is clearly recognized by the speech recognizer should be included in the summary and use confidence scores along with information retrieval metrics to find the most salient fragments of speech. Research at CMU on meeting summarization [61] is closest to the work...
we propose here; again, their interest is in dialog feature identification. They focus on identification and removal of disfluencies, linking of question/answer pairs, and identification of sentence boundaries. Summarization is achieved by extracting a small number of turns from the dialog, editing the turns to remove disfluencies and including answers that are linked to the extracted turns, to form the summary. This is quite different from the approach we propose where extracted information will be transformed in a variety of ways to produce many perspectives of the meeting. A variety of extracted information (e.g., terms, linked terms, discourse topics, points of agreement and disagreement) will be merged through information fusion and language generation processes to create different levels of abstraction, as well as coherent text, that provide an overview of the meeting.

4 Proposed Research

The proposed work has three thrusts: automatically extracting data from meetings and organizing it into hierarchically structured “maps”; using the information from such a map to create different types of summaries; and developing dynamic techniques for map construction. This section describes these three efforts, followed by an evaluation plan for assessing the contributions in each area.

4.1 Building Meeting Maps

The foundation for the meeting maps are the tools that we will develop for analyzing speech data and automatically labeling it in terms of the word transcription, speaker identity and topic structure. While we will build on existing work, there are important problems that arise with the meeting data that will require additional research in this project, as described below.

4.1.1 Transcription Technology

Speech recognition with error modeling Even with significant advances in ASR performance there will be errors in the transcription that could be compounded in later language processing. As we have seen in our previous work on information extraction [33, 34], language processing is improved by explicitly modeling these errors, i.e. missing information is less harmful than incorrect information. Here, we will include generic “error” as part of a word lattice generated by a recognizer. The error token will be associated with the probability mass of the set of low probability hypothesized words in a parallel position in the larger parent lattice generated by the speech recognizer. Using the confusion network representation of [28], the process of adding error tokens becomes one of lattice size reduction, replacing sets of low probability word arcs with an error token. This representation also has the advantage of keeping the word lattice small for more efficient language processing in subsequent analysis, and to support accurate word confidence estimation [11, 45]. The confidence scores provide “weights” for the different word alternatives that can be used in subsequent processing including passage retrieval. The use of error tokens in the output of a first pass of recognition can also lead to efficient multi-pass processing in the case of dynamic meeting map generation (see Section 4.3).

Utterance and topic segmentation using prosody Because speech recognizer output does not include punctuation, utterance boundaries are “hidden” in the stream of words. Continuous speech in meetings must be broken into coherent utterance units to enable automatic dialog structure annotation [46], as well as for other higher-level tasks including summarization. Unlike text, speech is rich in prosodic cues (such as pauses, boundary tones, pitch resets, and lengthening of syllables) that signal various types of meaning-related boundaries. We will develop techniques that make effective use of prosody to find boundaries of both sentence and topic units in the meeting data. In past work we have demonstrated marked improvements on these tasks in more controlled domains, namely news broadcasts and two-party phone-conversations [40, 51], showing that prosodic cues can lower the detection error rate by over 25% for topic boundaries and close to 20% for sentence boundaries. The detection of disfluencies is an integral component in the segmentation model, as disfluencies often occur around segment boundaries and use similar prosodic cues [48].
Figure 3: Spectrograms of 6 channels from a meeting excerpt showing some common signal features and a particular pattern of speaker turns.

Our approach employs features directly extracted from signal measurements (e.g., pitch tracks) and automatic speech recognizer output (e.g., durations) without error-prone and labor-intensive hand-labeling of phonetic prosodic categories (such as pitch accents and boundary tones). Features are selected automatically according to whether they help a given task. For certain features, such as fundamental frequency, we have also developed sophisticated modeling and smoothing techniques to increase robustness in the face of inherently errorful measurements [44]. We have developed effective ways to combine prosodic information with statistical language models of boundary types, using hidden Markov models and decision trees to make optimal classifications based on the combined evidence. The classifier components can be trained separately to make efficient use of the data [40, 51].

Meeting data will be a new challenge for these techniques. Some of the features that were relevant in Broadcast News and phone conversations should also apply to meeting data. On the other hand, we expect meetings to exhibit unique characteristics and new opportunities for prosodic modeling. In particular, speaker overlap has to be dealt with head-on, since strict acoustic separation of speakers is not possible. The timing of overlaps and interrupts across speakers will have to be integrated into the model for segmentation. For example, preliminary results show that sentence boundaries, pauses, and disfluencies, all of which feature prominently in our segmentation models, are also good cues to “points of opportunity” for interrupts by other speakers [42].

Patterns of speaker activity Speaker turns—i.e., who speaks when and for how long—provide rich information about the overall meeting type, the phases and events within a meeting, and the roles of different participants. Figure 3 shows a 40 second excerpt from our existing data collection. Each row shows the auditory spectrogram of one microphone’s channel. The segment begins with speaker A talking. At $t = 132$s, speaker B takes over, however speaker A regains the floor at $t = 152$s. From this pattern, we might infer that the discussion is relatively informal, and we can deduce a certain amount about the relationships between speakers A, B and C based on how prepared they are to interrupt one another. In this figure, we can see several of the difficulties that emerge when we try to recover speaker turns from these signals, including: (1) Crosstalk: Although each participant’s voice is shown most clearly in their ‘own’ channel, attenuated versions of the signal appear in most of the other channels, due to the weak but
significant coupling between the speaker's voice and every microphone in the room. (2) Overlap: The widely-observed overlap between speakers further complicates speaker activity detection since we cannot use simple heuristics such as assuming that the single highest-energy channel is the only active speaker. (3) Breath noise: A poorly positioned microphone will pick up the sound of a participant's exhalations. This breath noise can easily have as much energy as the participant's voice, although with a characteristic low-frequency spectrum, and without any crosstalk to other channels. (4) Nonspeech: Meetings include a variety of nonspeech events such as chair sounds, and objects being moved on the conference table, that do not derive from any of the participants.

Certain classes of 'interesting' nonspeech sounds—such as laughter—may benefit from the information available in cross-channel analysis. Although laughter has distinctive acoustic characteristics, it may be its social characteristics—i.e., the way it erupts simultaneously across several participants in a group discussion—that will provide the most salient features for deducing the state and nature of the meeting.

Cross-cancellation and position estimation Crosstalk is a general problem, and we aim to cancel it out by subtracting appropriately filtered versions of the corresponding close-mic channel. The Block Least Squares algorithm [59] is able to make an accurate estimate of the linear coupling between channels [7], but its performance is limited in the meeting data due to participant movements, which are quite common. We will pursue fast and noncausal solutions, based on common patterns of coupling dynamics as observed in the recordings.

Coupling variability has, however, a useful aspect: when the coupling changes, we know that movement has occurred. By comparing the changes between the various microphones, we can decide if it is the speaker or the listener who has moved. The net result is that crosstalk turns the microphones into passive triangulation device on the head of each participant. We will investigate the usefulness of this motion information within the meeting maps.

4.1.2 Mapping Content

As outlined earlier, meeting maps fall broadly into two categories: those focussed on content, and those related to interaction among the participants. In this and the next section we give examples of each of these categories. Content analysis will make use of some of the techniques developed for textual natural language, though we expect interesting research problems to arise from the interactive, spontaneously shaped structure of meetings which no doubt violate many of the assumptions underlying text-based methods. Also, we will focus our attention on the use of acoustic (especially prosodic) features for content-extraction tasks.

Information extraction. In order to describe the semantic content of a meeting, it is useful to identify key phrases and phrase types known a priori to be important for summarization, such as named entities (persons, places, organizations), dates, and the like, as well as word sequences such as “action item”, “in summary”, etc. A first step in the research will be to determine the set of the most important phrase types, recognizing that there will be more than one way to say each type because of language variability as well as ASR errors. Next, we will extend our previous work on statistical phrase language models for extracting such information [33, 34], to operate on word lattices and account for disfluencies. This effort will also build on our previous work on disfluencies in language modeling [47, 41, 43]. In addition, we propose to use cues from utterance, topic and speaker segmentations in the phrase decoding. The extended tools will also be leveraged to provide shallow parses for use in the summarization stage.

Certain important terms will not fall into previously identified categories, such as the list of educational background categories in the example in Section 2. We will use a combination of factors—such as word confidence, prosodic prominence, frequency and position in the meeting segment, etc. – to classify content words and noun phrases as being candidate “terms”. Next, we will develop clustering and word association techniques to determine groups of words/phrases that are associated either because they refer to the same entity or belong to the same class. These groups are needed to resolve pronominalization and anaphora for summarization, but also to reduce the number of “terms” that might be included in an outline-style display. Particular emphasis will be given to the integration of multiple cues (lexical, syntactic, prosodic, and speaker-related information) to determine these groups. Clustering to determine semantic groupings (such as in the list example) might also benefit from the use of the Wordnet semantic hierarchy.
**Information salience.** An important issue for information retrieval and summarization is identification of salient points of the meeting, or more generally labeling utterances in terms of the relative utility of the information. Multiple dimensions must be considered here, including relevance to the meeting topic and level of acceptance by other participants. Low-salience information can be filtered before summarization, while high-salience regions (e.g., action items, votes, announcements) will be preserved. A first step in the research will be identifying the different dimensions of “saliency”, and then combining prosodic, lexical and non-lexical cues to predict a running salience “confidence score”. Lexical measures of salience are typically based on information theoretic measures (such as entropy [14]), whereas acoustic-prosodic cues would correlate with indicators of interest or involvement from the participants, such as higher than usual speech volume, speaker overlap, and pitch excursions. The meeting environment will provide a rich testing ground for both new and old measures of salience.

**Topic identification and tracking.** We will use a combination of lexical and prosodic cues to identify locations of topic changes, after which we can apply methods to identify and track recurring topics in disjoint segments [1]. Recurring topics can be identified by a pairwise statistical comparison of the word-distributions of all the topic segments; similar word-distributions (e.g., as measured by a cosine-distance or symmetrized relative entropy measure) indicate similar topics. Topic sections can then be summarized individually or as a whole, as discussed in Section 4.2. Finally, to provide meaningful, succinct topic identifiers for human consumption, we will generate labels from the data by extracting salient phrases, e.g., those that occur near the introduction of a topic, consist of frequently occurring content words, and/or are prosodically prominent.

**4.1.3 Mapping Interactions**

Meetings are normally arranged for the purposes of getting things done. An interactional map of a meeting would involve such things as how the speakers structure and coordinate their turns, which aspects of their own identity and their relationships with other participants they choose to emphasize or downplay, and the prosodic, syntactic, and pragmatic aspects which are brought to bear in accomplishing things in a meeting. The analyses described in this section are a bridge between the representations presented in Figures 2 and 3.

**Dialog structure labeling.** Verbal interaction occurs at four main levels: propositions or content, actions accomplished by speech, participation management by turns, and social and affective marking. Even interactions focused on information exchange typically are organized by turns that display or challenge alignment – for example, by sustaining the topic with a new contribution, by acknowledgement, by demonstrating understanding, by paraphrase, or by repetition. Meeting maps can concern each of these four levels and their interaction, depending on the users’ goals [24].

A further level of representation and analysis for meeting data is the labeling of individual utterances as statements, questions, acknowledgments (“I see”), and so on. Such dialog act labeling is the first level of analysis in the identification of dialog structure, which can include higher-level units, such as question-answer pairs or digressive discussions. Dialog structure labeling is an important tool in support of tasks such as summarization. For example, in a concise summary certain types of utterances (such as acknowledgements and agreement) can safely be ignored, whereas others (such as questions and statements) are much more important. Dialog acts also convey sociolinguistic information: an analysis of who tends to issue statements rather than acknowledge or agree gives insight into the power dynamics of meeting participants. Acts that commonly occur in meetings include greetings, address, information requests, directives or requests for action [9, 10] and assessments. These often carry affective or social markings by variations in how they are accomplished, by laughter, by vocal information, by names, and by other discourse markers [13]. Automatic dialog act labeling will be based upon the word content of utterances, their prosodic features, and the surrounding dialog acts [46]. While dialog act labeling is a well-established task in computational linguistics, meeting data poses many new challenges, such as the inclusion of more than two participants and a large number of overlapping utterances. Furthermore, we propose to investigate unsupervised dialog act labeling (using automatically induced classes) to identify higher-level categories of meetings (such as arguments or “hot spots”), thereby circumventing the need for expensive human labeling of dialog acts as well as allowing for the potential discovery of new labels.
Overlaps. Due to the multiple interests and pressures which speakers must balance at any given point, meetings are often very complex. A case in point is overlaps: In contrast to the idealized view of conversations as proceeding in an ordering fashion from one speaker to the next, meetings are often full of overlaps. Furthermore, overlaps serve very different functions, from encouraging a speaker to continue (“mm-hmm”), to trying to usurp the conversational floor.

In a preliminary examination of overlaps in a 45-minute meeting of 4 people, 215 overlaps were detected. Overlaps occurred during almost every minute of the conversational speech session, and seemed to be about as frequent during the opening minutes of the meeting as during the rest of the meeting. The maximum number of overlaps seemed to occur during the contentful part of the meeting, which was presumably the part when people were most involved. Most overlaps involved only 2 speakers, but a small percentage involved more.

In our proposed work, these overlaps will be more finely subdivided according to the following categories (and others): backchannels (of various types), anticipating the end of a question and answering early, finishing someone else’s sentence for them, accidental interruptions (misjudging “transition relevance points”), and usurping the floor. In addition to distributional analyses, we will examine how the different categories of overlaps function and which ones correlate with other variables, such as consensus or meeting outcomes. The literature in linguistics and conversation analysis [15, 25, 12] indicates that these kinds of distinctions have identifiable syntactic and prosodic characteristics. Such characteristics will be leveraged in automatic detection work.

Consensus vs. disagreements. An important distinction in meetings is whether interlocutors establish consensus or disagree, and which members of the meeting adhere to the different viewpoints. These categories, apart from being useful for summarization purposes, can also point to important sections of a meeting; for example, disagreements tend to expose underlying reasoning and may be useful for later examination. Researchers in conversation analysis have claimed that agreements and disagreements can be identified by form-based cues (e.g., that disagreements, being “dispreferred”, are preceded by longer pauses, discourse markers such as “well”, and tend to be more verbose [27, Chapter 6][36]). We will test these and other cues (e.g., energy distribution within and across turns) against our meeting data, and model them statistically for the purpose of automatically gauging the level and timing of consensus in a meeting segment. A variation of this approach may be useful for identifying decision points reached without consensus.

Conversational style. Sociolinguistics and discourse research are relevant for the identification of “trouble spots” in communication. Among people from the same cultural background, turn taking proceeds in a smooth and regular way. Communication difficulties can be signalled by something as subtle as intra-turn pauses which are either shorter or longer than expected.

Where speakers come from different cultural backgrounds, the norms of one group may conflict with those of the other. For example, in Tannen’s study of conversational style [50], she contrasted the “high-involvement” style of a person from New York with the more relaxed expectations of someone from Los Angeles. High involvement style is characterized by a tendency to ask a lot of questions, and to overlap animatedly with the other speaker. The Los Angeles speaker perceived the New Yorker as being nosey and overly aggressive, whereas the New York speaker perceived the Angeleno as being aloof and evasive.

Cultures also differ in self-presentation, and structuring of information. Within the context of a housing (or job) interview [16] or a school counseling session [8], such differences can greatly affect the applicant’s opportunities. Within the context of a meeting, these differences may lead to slower progress toward goals, greater frustration, and more ”trouble spots”.

Including conversational style will enhance the resolution of interaction maps, enabling for example discrimination between true trouble spots and simply stretches involving high involvement style.

4.2 Using Meeting Maps: Summarization

Meeting summarization will be a key application for the meeting maps’ data. The term ‘summarization’ is used here to include several types of information extraction at varying levels of granularity, ranging from indicative summarization (e.g. a list of keywords indicating the topics addressed in the meeting), gisting (a list of phrases providing the most
relevant factual information relating to each topic discussed), to a complete summary of the meeting in the form of a coherent text. In addition to level of detail, the summaries may vary depending on the characterization of the meeting itself. For example, a meeting involving discussion, dissension, subsequent resolution of dissension and decision making could be summarized quite differently than a software design review meeting. Thus, the first part of our work will include a characterization of meeting types, summarization types, and summarization strategies for the different typologies.

To illustrate the research needed for the summarization technology, consider the task of generating a fluent, informative summary of an argumentative meeting. Most current single document summarization systems work by extracting a set of key sentences and stringing them together to form the summary [23, 26]. For speech, this would result in a sequence of disconnected statements by different participants, something that is very unlikely to be a useful summary. Instead, as illustrated in Section 2, we might want an informative summary that described the different viewpoints taken, which speakers aligned with each viewpoint, and the final decision made. This could be represented as a summary "plan", with three slots (viewpoints, partisans, decision) dictating summary content. Research would be needed to determine how to identify and extract content for each slot, how to merge different descriptions of the same fact (e.g., a viewpoint) into a single presentation, and finally, how to transform the wording of the transcript (that is, spontaneous speech with disfluencies) into a coherent, fluent summary.

To see how this might work, consider the example informative summary from Section 2. Input to the summary generator would include information extracted from the input speech as represented in the meeting map. Critical features for generating this type of summary include topic boundaries (in order to identify the beginning of the discussion of survey forms), terms characterizing the first topic (in order to produce a description of the segment topic), locations in the transcript where disagreement and agreement occurred, identification of the speakers who expressed disagreement and agreement, identification of speech acts for each turn in the dialog and segmentation into sentences or phrases that convey the exact points of disagreement.

Given this information, we can see how the system might produce the final sentence of the example summary (Finally, they agree on the categories undergrad, graduate student, post-doc, professor, and a last choice with the words visiting post-Ph.D. researcher). If the first turn in the transcript segment shown in Figure 1 is labelled as a bid for agreement, inferred from the phrase maybe they’re fine... and the first two turns are also labelled as question (from repeat the categories you had) and answer, then the summary generator can decide to include mention of this agreement in the summary and link the reference categories with the list undergrad, grad, post-doc, professor, other to create the first part of that sentence.

Summarization technology will be developed to extract noun and verb phrases (from the content map) within the identified important turns of the dialog (from the interactions map), and piece them together with other constructed phrases referring to dialog structure and the act of agreement (e.g., Finally, they agree...) to produce the full summary sentence. We can use the information fusion techniques we have developed to identify and merge similarities across documents in multi-document summarization [30, 2] to merge phrases representing similarities in viewpoint. There are many open questions and problems here and we will need to vary our target for summary output as we determine what information the technology is able to extract; for example, it may be difficult to determine that post-Ph.D researcher is the last choice, but we certainly can extract the term to include in the summary.

This top-down approach to summarization can be used for informative summaries, with the plan guiding selection of content. It is similar to domain-dependent approaches [37, 57] that produce multi-document summaries of specific kinds of news stories (e.g., terrorist events, or natural disaster stories) by generating sentences representing the particular template slots (e.g., event type, perpetrator, victims, time, location) derived by an information extraction system [6]. A summary plan can be developed based on meeting genre. It is in contrast, however, to most current domain independent approaches which are bottom-up, with content opportunistically emerging from clustering to find similarities [17, 5, 60] or from statistical metrics that find important sentences in the text [4, 26, 49]. We will explore the use of bottom-up strategies to produce the indicative summaries describing meeting topics (see Section 2), using identified topic segments and associated terms and phrases.
4.3 Dynamic Meeting Maps

Much of the work described in Section 4.1 assumes fairly sophisticated automatic transcription technology, for which the processing would likely be offline and therefore result in static pre-computed meeting maps. However, there are important reasons to consider “dynamic” maps, formed at the time of a user’s interaction with the system. For instance, it may be the case that a user cannot afford offline processing, because they need answers immediately. Or, it may be the case that the set of pre-existing maps do not contain the meeting information most relevant to a particular user, simply because the items of interest were not in the ASR system vocabulary used at the time of off-line processing. In other cases, a given map might contain more information about a meeting than a user is interested in. In this case, a reduced resolution of the map, while being more general, might not be the best representation—resolution would be reduced, but without highlighting the points of interest to the user. Furthermore, computational and storage limitations might well make it practically infeasible to precompute all possible maps in advance.

In order to address these problems, we propose two steps towards generating dynamic meeting maps. In the first case, we look at query-dependent ways to generate meeting maps. In the second case, we look at progressive recognition strategies aimed at on-line processing.

4.3.1 Query-Dependent Meeting Maps

One way to improve the usefulness of meeting maps is to base them on user-provided queries, where the map is derived dynamically from a specific written or spoken phrase about meetings. The map is then generated or modified at the time of query, such as when a user is attempting to extract information from a pre-recorded set of meetings. The approach is to modify the input given to the map generation procedures, described above, to contain only that information most relevant to the query. By semantically analyzing the query, content words and phrases possibly related to the query and contained within the vocabulary of the meeting recognizer will be flagged as relevant. As in Section 4.1.2, this analysis would benefit from the use of relations contained within the Wordnet database. The sensitivity of relevance (i.e., the neighborhood within which words are considered relevant to a query) can be widened or narrowed by adjusting a word similarity sensitivity parameter. The resulting set of relevant words can be used to filter the existing static maps, highlighting regions of greatest interest. Alternatively, we can construct query-specific input to the map generation procedures themselves.

In order to generate query-specific maps, modified language models can be constructed by reducing the vocabulary of the recognizer to contain only those words that have been flagged as relevant. All other words can be collapsed into a single (or a small number of) query-irrelevant word clusters, each modeled as a single vocabulary item. Such a model will prefer those words and phrases most relevant to the query; everything else will be recognized as irrelevant and omitted. The language model can be used to rescore the confusion network representation of the speech, either with or without acoustic model rescoring (in case new words were introduced into the vocabulary). The updated ASR output would be provided to the map building procedures, resulting in query-dependent meeting maps. (Since the processing steps above the level of the word will be much less costly than the word transcription step, it will be possible to do these additional stages of processing relatively rapidly.)

4.3.2 Coupling ASR and Information Extraction

In order to enhance ASR performance for the purpose of meeting map generation, we will investigate a cyclically coupled, multi-stage process of speech recognition and information extraction. At each iteration of this process, analysis results are exchanged between the recognition and information extraction modules to enable the subsequent use of gradually more refined models. At the first stage, simple, “shallow” recognition strategies like keyword or phrase spotting can be employed to identify regions which might potentially be relevant for the desired meeting map. These key-phrase spotters may focus on general, topic-independent structural markers like “let’s move on”, “let’s start”, “final point”, etc., or they may take into account specific user queries about the meeting (as described above) in which case they detect topic-related content words. Key-phrase detection can be combined with the output of the topic segmentation and non-lexical event modeling modules described above, which provide information about widely-used non-lexical discourse markers such as intonational phrase onsets, throat clearing, etc. These different information sources are combined to define broad regions of interest which are then selected as input to a speech recognizer.
using a larger recognition lexicon and possibly more sophisticated (e.g. speaker-adapted) acoustic models. The output generated by this recognizer will consist of word transcriptions of the previously identified utterances or parts of utterances. At this point, further refinement techniques such as statistical error correction can be integrated, which renders the basic recognition output more reliable and more suitable for subsequent processing such as summarization. The error correction process is likely to identify useful information not available to the basic recognition system, such as proper names which are not included in the recognition lexicon. Similarly, the subsequent summarization process will generate additional expansions of the recognized word string, such as synonyms. These and other analysis results from higher-level modules can be re-integrated into the recognition system in order to improve recognition performance on later parts of the meeting and on future meeting recordings. Proper names and synonyms identified by the information extraction/summarization modules can be associated with automatically constructed pronunciation models and added to the recognition lexicon. Furthermore, the language model of the recognizer can be improved by including n-gram probabilities for these new items, which can be obtained by interpolating their predicted unigram probabilities with more general class-based language model probabilities.

This hierarchical analysis strategy, where general salient regions are identified first using simple models and are then decoded in detail using more refined models, places particular emphasis on speed of analysis and on the detection of the most salient points in a meeting, as opposed to a fine-grained, exhaustive analysis of the entire meeting structure. The development of such a strategy is therefore a necessary prerequisite for the potential real-time application of a meeting browser.

### 4.4 Evaluation

Meeting maps are built upon a foundation of technologies for recognizing speakers, words, topics, etc. in speech. Each of these recognition tasks can be evaluated independently by comparing recognized structure to hand-labeled test data. For speaker and word recognition, standard accuracy measures will be used to assess performance. For tasks such as topic detection and information extraction, the standard recall, precision and F-measure (harmonic mean of precision and recall) scores will be used.

Three main methodologies have emerged for summary evaluation [19]: task evaluations [21], comparison against an “ideal” summary (see the upcoming NIST evaluation of summaries [32], and subjective judgments by humans of the generated summary [57]. For the proposed research, we will use minutes of formal meetings as an ideal representation of summary content. These minutes have been generated by meeting participants for their own use (independent of this project) for several of the meetings recorded in our data collection effort. We will score a generated summary by comparing the number of discussion and decision points against those in the meeting minutes. To do this, we will break down the generated summary into clauses, using either manual annotation of output or automated chunking by modifying the summary generator. A human judge will compare each generated clause against the minutes to determine if they overlap in content (i.e., identical phrasing is not necessary). For an argumentative meeting, for example, this methodology will allow us to determine if the automatically generated summary contains the most important points of disagreement and decisions made. While evaluation of summary content will be a critical measure of quality, we will also experiment with different methods for judging coherency, relying on human subjective judgments of the generated summary. We will also explore different methods for task evaluation. For example, we will evaluate the indicative summaries that overview the different topics discussed through an information retrieval task, measuring whether the summary allows a user to determine meeting relevance more quickly and more accurately than other representations of the meeting (e.g., the full transcript, the index terms alone, etc.)

We plan to evaluate the research on dynamic meeting maps using similar paradigms as above (e.g. both task-level evaluation and accuracy of transcription and information extraction modules), comparing performance with and without customization to specific queries. In addition to meeting relevance, we will look at segment relevance, using queries designed to relate only to specific segments of a meeting.
5 Preliminary Work

A prerequisite for the proposed work is the availability of large amounts of realistic data. We have begun work in this area by instrumenting and using a room at ICSI for multichannel recordings of meetings. As of this writing we have recorded roughly 45 hours of natural meetings (i.e., meetings that would have occurred whether or not they were recorded). We have transcribed about 30% of these data, from which the introductory examples were taken. Note that there are separate microphones for each participant in addition to 6 far-field microphones, and there can be as many as 16 channels. Consequently the sound files comprise hundreds of hours of recorded audio. The total number of participants in all meetings so far is 237, with 49 unique speakers. An effort is under way to build a portable collection system at UW, and human-generated minutes will be collected for these meetings. By the time this project would start, we expect to have roughly 100 hours of data collected that could be used in the research.

We have developed a baseline automatic transcription system by adapting the SRI large-vocabulary conversational speech recognizer [45]. We found word recognition accuracy on close-talking microphones to be on par with that on standard phone conversation corpora (for which the system had been developed), even without special adaptation of the acoustic or language model to the meeting domain. This result is important as it gives us confidence in the portability of our speech models, and allows us to use the recognizer for automatic alignment of word transcripts and other tasks. At the same time, we learned that speaker segmentation and recognition from non-close-talking microphones will be hard research problems in this domain [31].

We have also studied overlap in four types of conversational speech corpora. Results show that both meetings and telephone conversations have high rates of overlap, suggesting that overlap is an inherent characteristic that should not be ignored in computational models of conversation. Results on word error rates using the same speech recognition system on each of the four corpora reveal that in regions of no overlap, recognition on meeting speech is at about the same level of accuracy as that for telephone conversations. Recognition does suffer from overlaps, even on close-talking microphone data. The errors are in the form of insertions, which should be partially addressable by cross-cancellation techniques, but which present an important challenge for further research. Finally, interrupts do not just occur in random locations, but rather are associated with hidden events (such as disfluencies and discourse markers) in the foreground speech. Interrupts tend to start after such events, suggesting the value of an integrated acoustic/language model for speaker segmentation in natural conversation [42].

We have implemented a speech activity detection system in order to pre-mark each speaker’s hypothesized segments for the human transcribers [35]: an HMM-based speech activity detector is applied to each channel, using a combination of standard cepstral features, and log-energy features normalized by the average across all channels in an effort to factor the source level out of the gain estimates. Speech activity ‘false alarms’ due to crosstalk are removed in a post-processing step that measures the correlation between simultaneously-active channels, removing some 40% of the errors.

In addition to this specifically related work, the project will be able to leverage software and expertise of the participants who have worked on core problems that will be needed for building the maps, including speech recognition, speaker ID [53], utterance and topic segmentation [40], information extraction from speech [33, 34], and summarization [37, 20, 29].

6 Project Integration

As is required for an undertaking of this scope, the team of investigators brings a range of experience and skills to bear on the problem. We have assembled a team of 12 senior researchers from engineering and social science disciplines at 4 research institutions. Together, we represent expertise in corpus linguistics, discourse analysis, social psychology, prosodic analysis, disfluency analysis, cognitive modeling, text summarization, information extraction, sound content analysis, speech transcription technology (including robust speech processing), and statistical modeling of spoken language.

Strong factors already exist that will guarantee a truly active collaboration among the institutions participating in this proposal. These four institutions (ICSI, SRI, U. Washington and Columbia U.) are already closely linked in other collaborative projects. Nearly all of the senior personnel involved have ongoing projects with one or more of
the other sites and we already have established a schedule of regular meetings which would be expanded to include an annual workshop for all participants in the proposed project. The proposed budget will also provide for frequent travel between sites for key partnerships within the project, and particularly for students at Berkeley, Washington, and Columbia to expand their breadth of understanding.

Overall, the proposed budget will permit the training of 10 graduate students in the range of topics required for this project. The methods taught will cut across traditional boundaries (speech recognition, prosodic analysis, text summarization, etc.) and so will provide the students with a broader education. The students that are so trained will facilitate the project integration since they will ultimately understand the cross-topic research better than their advisors. In addition, the faculty involved in this research will work to integrate the effort into the curriculum by way of class projects and the analysis of course discussions using the techniques described here. Finally, we propose to develop short courses at the different sites, which the participating students from all sites would attend. The short courses could also be made open to students from other universities.

7 Impact

The research agenda proposed here is highly ambitious and risky in terms of ultimate success—many of the tasks are still poorly understood pending careful study of the actual meeting data, which is novel and has not been the focus of much attention in speech and natural language processing to date. However, even if only partially successful, the theories and algorithms developed will fundamentally change the way we can leverage information technology for human-human interaction.

The proposed project will also provide key scientific insights into human interaction in the natural conversations of purposeful meetings. The systematic analysis of prosodic and other vocal information will allow the study of the relation between features at different levels of interactional discourse, such as vocal marking of topic change, vocal features indicating agreement and disagreement, and marking of shifts between statements of information and proposals for action. In addition, the computational tools that will be developed will be useful for further study of interaction in meetings by other scientists.

The endorsement of this project by several key relevant companies underscores the potential utility of this technology for practical applications as well. Aside from understanding face-to-face meetings, the tools that we develop could also be used for information mining of call center conversations, an area of key commercial importance. Advances made in the project could facilitate access to information from the large amount of publically available audio from government meetings such as Congressional hearings. In all of these cases, the methods and systems developed in this project will be made available to any who wish to use it.

The data collected and annotated (both by human annotators and automatic tools) will provide a novel data resource to the speech, natural language processing, and social sciences communities, and will no doubt engender new and diverse research at other sites. The database we will develop will be unique in its nature (natural multi-party meetings), relatively large size (more than 100 hours of speech), and rich annotation at multiple levels of analysis, all the way from low-level acoustic-prosodic features to high-level semantic and pragmatic categories.

The work is being done by a collaborative group including 8 faculty and 10 graduate students at 3 academic institutions (Berkeley, U Washington, and Columbia). As noted in the previous section, the project will not only provide a research direction for these students, but will enhance understanding of the use of computer tools to aid the utility of human conversation, a social goal that is typically lacking in engineering curricula. Aspects of this research will be infused in coursework at the 3 institutions.

Six of the twelve senior investigators on the project are women. We expect that this gender-balance in a heavily engineering-oriented project will provide positive role models and incentives for recruiting female students to the research fields represented, in many of which women are historically under-represented.

---

2ICSI is closely affiliated with UC Berkeley, and the PI is a Professor on the EE faculty; the students he and his ICSI and UCB colleagues will supervise are Berkeley graduate student in EE, CS, and Linguistics.

3See the accompanying endorsement letters from two speech technology companies (Nuance and VoiceSignal) and from a high profile business that incorporates voice technology (Visa International).
8 Previous Work

Three of the principal investigators have had NSF support in the past five years: Nelson Morgan, Mari Ostendorf and Andreas Stolcke.

NSF IIS-9712579 ($784,518)
Title: Automatic Speech Recognition Based on Syllable-length Acoustic Models
Award Period: 09/01/97–08/31/00
PI: Nelson Morgan

This project investigated the use in ASR of units with a longer time basis than the traditional phones or sub-phone units. The view was that such units might be a better match to fluent speech, for which pronunciations often deviated from canonical forms. New signal processing methods for robust speech recognition were also developed. Four Berkeley Ph.D. dissertations came out of this work, as well as three journal papers [Speech Communication, two in (29) 2-4, and one in (31) 1], and 10 conference papers. The acoustic processing methods proposed here will incorporate many of the methods discovered in the course of this NSF-funded project.

NSF IRI-9618926/9996450 ($451,833, including REU supplement)
Title: Modeling Structure in Speech above the Segment for Spontaneous Speech Recognition
Award Period: 03/01/97–06/30/01 (including no-cost extension)
PI: Mari Ostendorf

This project looked at the use of different levels of linguistic structure in acoustic modeling, including: use of syllable and prosodic structure in clustering observation models and topologies, dynamic pronunciation modeling as a function of discourse and prosodic context, pronunciation modeling in terms of linguistic features, and dependence modeling for adaptation. Two Ph.D. students supported by this project are expected to graduate June 01 and March 02. Two MS students and a research associate (Katrin Kirchhoff, now a Research Professor at UW) were also supported in part by this work. Two journal publications associated with this work have appeared (IEEE Trans. SP, 48(6) 2000 and Phil. Trans. Royal Society 358(1769) 2000); one paper is in review and another two are in preparation. Several conference papers have appeared or are in preparation. The work proposed here will be able to take advantage of some of the speech recognition infrastructure at UW associated with this project and other previous NSF-funded work.

NSF IRI-9619921 ($769,968)
Title: Modeling and Automatic Labeling of Hidden Word-Level Events in Spontaneous Speech
Award Period: 03/01/97–02/28/02 (including no-cost extension)
PI: Elizabeth Shriberg, Co-PI: Andreas Stolcke

This project investigated the statistical modeling and automatic detection of “hidden events” in speech recognition output, such as disfluencies, sentence boundaries and dialogue act types that are either absent in written language or explicitly marked by punctuation, but need to be recovered for effective speech understanding. Research under this funding demonstrated that prosodic information could be successfully leveraged for these tasks. The project has directly funded 7 graduate students (at Stanford, U.C. Berkeley, Boston U. and U. Bilkent), four of whom are women. It has also provided advisory and dissertation research resources support for 13 additional students at Stanford U., U.C. Berkeley, U. Colorado, U. Minnesota, MIT Media Lab, U. College Dublin, U. of Alberta, U. Stockholm, U. Venezuela. Discourse-labeled data supported by the work has been distributed by the TalkBank initiative (LDC); prosodic and disfluency databases have been distributed to researchers at U. Minnesota, MITRE, and Dragon. The project research has so far led to a patent application (pending), five journal publications: Language and Speech 41(3-4), Speech Communication 32(1-2), Computational Linguistics 26(3) and 27(1)), Journal of the International Phonetics Association (to appear); and numerous conference papers. The work proposed here will rely heavily on the expertise developed with this prior NSF funding, particularly prosodic modeling for speech segmentation, disfluency processing, and dialogue modeling.
References


NELSON H. MORGAN

EDUCATION:

1977  B.S.  (Electrical Engineering, with Highest Honors) UC Berkeley
1979  M.S.  (Electrical Engineering, NSF Fellow) UC Berkeley
1980  Ph.D. (Electrical Engineering, NSF Fellow) UC Berkeley

ACADEMIC/PROFESSIONAL APPOINTMENTS:

2000-  Professor in Residence, EECS Dept., UC Berkeley
1999-  Director, International Computer Science Institute
1992-1999  Adjunct Professor, EECS Dept., UC Berkeley
1986-1988  Vice President/Engineering, SAM Technology, Inc., San Francisco
1984-1988  Chief Engineer, EEG Systems Laboratory, San Francisco
1980-1984  Chief Researcher, Speech Research Laboratory, National Semiconductor
1981-1982  Lecturer, Engineering Department, San Francisco State University
1967-1977  Independent Audio Engineer

SELECTED PUBLICATIONS/PATENTS CLOSELY RELATED TO THIS PROPOSAL:


OTHER SELECTED PUBLICATIONS:


**SYNERGISTIC ACTIVITIES**

Co-developing and teaching (with Ben Gold) the Berkeley EECS graduate course in speech and audio signal processing. The book co-authored with Gold mentioned above is the textbook for the course.

Former co-editor-in-chief of Speech Communication, an international journal; still on Editorial Board.

Board Member, Applied Voice Input-Output Society

Scientific Advisory Board Member, IDIAP, Switzerland

**OTHER COLLABORATORS FROM THE LAST 48 MONTHS**

Other than those listed in the publications above, this list should include Steve Renals, Chuck Wooters, Jeff Bilmes, Brian Kingsbury, Michael Shire, Su-Lin Wu, Andreas Stolcke, Dan Jurafsky, Jean-Marc Boite, and Mari Ostendorf.

**GRAD STUDENTS AND POSTDOCS**

Grad students supervised:

Chuck Wooters, Yochai Konig, Grace Tong, Kristine Ma, Su-Lin Wu, Nikki Mirghafori, Eric Fosler-Lussier, Warner Warren, Jeff Bilmes, Brian Kingsbury, Michael Shire, Andy Hatch, Dan Gildea, Adam Janin, Barry Chen, Dave Gelbart (total of 16, out of whom 7 have completed their PhD's, two have left with MS's, 1 is nearly complete with his PhD, two are nearly done with their PhDs, and the other 4 are active PhD students).

Postdoctoral Fellows supervised:

Steve Renals, Gary Tajchman, Dan Jurafsky, Florian Schiel, Holger Schwenk, Javier Ferreiros (6)

**ADVISORS**

Nelson Morgan's thesis advisor was Robert Brodersen. There was no postdoctoral period.
Daniel P.W. Ellis
Dept. of Electrical Engineering, Columbia University
500 W. 120th St. Room 1312, New York, NY 10027
tel: (212) 854-8928 email: dpwe@ee.columbia.edu
fax: (212) 932-9421 http://www.ee.columbia.edu/~dpwe/

Professional preparation:
Cambridge University, U.K. Engineering / Elec. & Info. Sci. BA(hons), 1987
M.I.T., Cambridge MA Electrical Engineering Ph.D., 1996
ICSI Berkeley CA Speech recognition 1996-2000

Appointments:
Assistant Professor, Dept. of Electrical Engineering, Columbia University August 2000-
Research into sound content analysis, understanding and retrieval including speech recognition
and source separation. Courses taught: Digital Signal Processing, Speech and Audio Processing
and Recognition.
Postdoctoral Scholar, International Computer Science Institute May 1996-April 1999
Speech recognition research including feature design, system architectures, information
retrieval and applications of hearing models and auditory scene analysis. With Prof. Nelson
Morgan, supervision of UC Berkeley graduate students in these areas.
Part of Machine Listening Group, researching computer models of human sound organization
including perception of ambient sounds and organization of music and other mixtures.
Tutor, MIT Office for Minority Education 1994-1995
One-on-one tutoring of undergraduates in areas of probability, signals and systems, etc.
Intern, Interval Research Corporation, Palo Alto CA June-August, 1994
Research on auditory representations as part of a speech separation project.
Member of Technical Staff, AWARE Inc., Cambridge MA 1991-1993
Research and development in high-quality audio compression, including contributions to the
MPEG Audio reference ‘committee code’ and a real-time engine for Macintosh computers.

Publications:
Related to this proposal:
D. Ellis (1999). “Using knowledge to organize sound: The prediction-driven approach to com-
putational auditory scene analysis, and its application to speech/nonspeech mixtures,” Speech
Communications 27, pp. 281-298.
D. Ellis & G. Williams (1999). “Speech/music discrimination based on posterior probability
Other publications:


Synergistic activities:

Since 1993, administrator of the “AUDITORY” list, an email discussion list for researchers in auditory organization (currently 587 participants in 32 countries).

Organizing committee member for Workshops on Computational Auditory Scene Analysis in 1997 and 1999; Treasurer of 1997 IEEE Audio Workshop; Co-chair of Workshop on Consistent and Reliable Acoustic Cues at Eurospeech 2001.

Reviewer for Speech Communications, Computer Speech and Language, IEEE Transactions on Speech and Audio Processing, IEEE International Conference on Acoustics, Speech and Signal Processing, Audio Engineering Society meetings, Neural and Information Processing Systems meetings etc.

NSF review panel member for 1999 ITR program.

Author of numerous public-domain software tools for sound analysis and processing, including co-ordination and maintenance of the “SPRACHcore” connectionist speech recognition package.

Collaborators & other affiliations:

Collaborators:


Advisors at graduate school:

B. Vercoe (MIT Media Lab), L. Braida (MIT EECS), B. Gold (MIT Lincoln Laboratory).

Advisees:

No formal role supervising graduate students, but close interaction with: J. Bilmes (U. Washington), E. Fosler-Lussier (Lucent), B. Kingsbury (IBM), N. Mirghafori (Nuance), G. Williams (Lernout & Hauspie), S. Wu (Nuance)
Curriculum Vitae

Katrin Kirchhoff
Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
215 EE/CS Building
Box 352500
Seattle, WA, 98195
katrin@isdl.ee.washington.edu

Research Interests
- speech recognition
- speech production, phonetics, linguistics
- pattern recognition/machine learning
- multilingual speech recognition
- dialogue modeling and speech understanding

Employment

University of Washington
7-2000 Research Assistant Professor, Department of Electrical Engineering
Seattle, USA

University of Washington
10/99-7/2000 Research Associate, Department of Electrical Engineering
Seattle, USA

Education

University of Bielefeld
8/99 PhD in Applied Computer Science
Bielefeld, Germany

International Computer Science Institute
10/97 - 7/98 Research student at the International Computer Science Institute
Berkeley, USA

University of Bielefeld
2/96 MA in English Linguistics/Applied Phonetics
Bielefeld, Germany

University of Edinburgh
10/91 - 6/92 Undergraduate Studies in Linguistics
Edinburgh, UK

Other Professional Activities
Curriculum development for graduate courses on speech and language processing at the University of Washington (with M. Ostendorf and J. Bilmes)
Program Committee member for Workshop on Human Language Technology 2001
NSF Panelist
Proposal Reviewer for the Science Foundation Ireland
Co-Organizer of the Workshop on Phonetic & Phonological Knowledge in Automatic Speech Recognition, Saarbruecken, Germany, 2000
Participant in the Research Workshop on Large-Vocabulary Conversational Speech Recognition, Johns-Hopkins University, Baltimore, 1997
Member of IEEE, ISCA and ACL
Reviewer for *Computer, Speech and Language, IEEE Transactions on Speech and Audio Processing*

**Selected Publications**


K. Kirchhoff. “Speech Analysis by Rule Extraction from Trained Artificial Neural Networks”, *International Conference on Spoken Language Processing 2000*, Beijing, October 2000


MARI OSTENDORF  
Professor  
Department of Electrical Engineering, University of Washington  
Box 352500, Seattle, WA 98195-2300  
mo@ee.washington.edu  
(206) 221-5748 (phone) (206) 543-3842 (fax)

SPECIALIZATION:
Statistical modeling for signal interpretation and generation, particularly speech and language processing. Current research efforts are in acoustic and language modeling for spontaneous speech recognition; prosody modeling for synthesis, recognition and dialog tracking; and information extraction from speech data.

EDUCATION:

Stanford University, Department of Electrical Engineering  
Bachelor of Science Degree, 1980  
Master of Science Degree, 1981  
Doctoral Degree, 1985

EMPLOYMENT:

1999 – present  
University of Washington, Seattle, WA 98195  
Professor, 1999 – present; Endowed Professor in System Design Methodologies.

1987 – 1999  
Boston University, Boston, MA 02215  
Professor, 1999; Associate Professor, 1993 – 1999; Assistant Professor, 1987 – 1993.

1995  
ATR Interpreting Telecommunications Laboratories, Kyoto, Japan  
Visiting Researcher.

1985 – 1986  
BBN Laboratories, Inc., Cambridge, MA 02138  
Scientist.

1981 – 1984  
Stanford University, Stanford, CA 94305  
Research Assistant.

Bell Telephone Laboratories, North Andover, MA 01845  
Member of Technical Staff.

PROFESSIONAL ACTIVITIES:


• Computer Speech and Language, US Editor, 1998 – present.

• Pattern Analysis and Applications Editorial Board, 1997 – present

• Professional Society Memberships: IEEE (Senior Member), ASA, SWE, Sigma Xi
SELECTED PUBLICATIONS:

Prof. Ostendorf has published over 100 papers on topics in speech processing; ten relevant papers are listed.


SYNERGISTIC ACTIVITIES:

- Co-Editor of the journal Computer, Speech and Language.
- Technical committee member for numerous workshops; chair of two workshops on speech synthesis and generation aimed at identifying infrastructure needed to forward the field.
- Member of the UW EE Strategic Planning committee and Undergraduate Curriculum Revision Committee.
- Supervised several undergraduate research projects, including 3 NSF REU students.
- Currently developing a 4-course speech and language processing curriculum for UW (with J. Bilmes), including graduate courses and a multi-disciplinary undergraduate course in speech interface design.

COLLABORATORS:

- PhD Thesis advisor: Robert Gray, Stanford University

  Thesis supervision in past 5 years: K. Ross (PhD 95), F. Richardson (MS 95), R. Bates (MS 96), O. Ronen (PhD 96), A. Kannan (PhD 97), C. Fordyce (MS 98), M. Siu (PhD 98), R. Iyer (PhD 98), M. Bacchiani (PhD 99), H. Shivakumar (MS 00), Y. Lobacheva (MS 00). (Total supervised: 10 PhD and 11 MS theses) Current students: I. Shaik, R. Bates, R. Fish, I. Bulyko, D. Palmer, S. Otterson, O. Cetin, C. Boulis.

- Recent collaborators: S. Shattuck-Hufnagel (MIT), A. Black (CMU), R. Sproat (AT&T), S. Chen (IBM), W. C. Karl (BU), David Castanon (BU), Hamid Nawab (BU), Nelson Morgan (ICSI), E. Shriberg (SRI), A. Stolcke (SRI), H. Gish (BBN-GTE), M. Meteer (BBN-GTE), J. Burger (MITRE), P. Price (Brava Bravo), G. Bernard (Boeing), and several EE faculty members at UW.
ANDREAS STOLCKE

Contact
Speech Technology and Research Laboratory
SRI International
c/o International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
Phone (510) 666-2969
Email: stolcke@speech.sri.com
URL: http://www.speech.sri.com/people/stolcke/

Professional Preparation

Technische Universität, Munich                        Computer Science      Diplom (M.S.), 1988
University of California, Berkeley                   Computer Science      Ph.D., 1994
          Dissertation: *Bayesian Learning of Probabilistic Language Models*
International Computer Science Institute, Berkeley, CA A.I. 1990-1994
          (Research Assistant and Postdoctoral Researcher)

Appointments
Senior Research Engineer, SRI International 1994-present
Visiting Researcher, International Computer Science Institute, Berkeley 2000-present

Research areas: speech recognition, statistical language modeling, computational linguistics, machine learning

Related Publications
(all available from http://www.speech.sri.com/people/stolcke/publications.html)


Other Publications


**Synergistic Activities**
Associate Editor, *IEEE Transactions on Speech and Audio Processing*
Co-Organizer, ACL-99 workshop on “Unsupervised Learning in Natural Language Processing”
Group Leader “Dependency Language Modeling”, 1996 Johns Hopkins Summer Speech Research Workshop

**Collaborators & Other Affiliations**
Horacio Franco (SRI International), Dilek Hakkani-Tür (Bilkent/AT&T Research), Daniel Jurafsky (CU Boulder), Andrew Kehler (UC San Diego), Christopher Manning (Stanford), Lidia Mangu (IBM Research), Nelson Morgan (ICSI, Berkeley), Mari Ostendorf (Univ. of Washington), Ananth Sankar (Nuance Communications), Elizabeth Shriberg (SRI International), Paul Taylor (Edinburgh), Gökhan Tür (Bilkent/AT&T Research), Mitchel Weintraub (Nuance Communications)
*Thesis advisor:* Jerome Feldman (UC Berkeley); *Graduate advisors:* Robert Wilensky (UC Berkeley), George Lakoff (UC Berkeley), Stephen Omohundro (ICSI Berkeley)
BIOGRAPHICAL SKETCH

Jeff A. Bilmes
Assistant Professor, Department of EE, University of Washington, Box 352500, Seattle, WA  98195-2500  
(206) 221-5236  Email address: bilmes@ee.washington.edu

Research Interests
Speech, language, and pattern recognition, Machine Learning, Information Theory, Graphical Models and Bayesian Networks, High-performance computing systems, Human-computer interfaces, Statistics, Data-mining, and Natural computation.

Education

Employment
University of Washington, Dept. of EE, Seattle, WA  
9/99 - Present  Assistant Professor.  
University of California, Berkeley, Dept. of CS, Berkeley, CA  
10/93 - 9/99  Research Assistant.  
International Computer Science Institute, Berkeley, CA  
10/88 – 9/91  Research Assistant.  
Massachusetts Institute of Technology, MIT Media Laboratory, Cambridge, MA  
10/91 - 9/93  Research Assistant.  
International Computer Science Institute, Berkeley, CA  

Awards & Professional Activities
2001 NSF CAREER Award
1998 U.C. Berkeley Samuel Silver Memorial Scholarship Award
Associate Editor, IEEE Transactions on Multimedia, 2000 -.
Session Chair, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999
Session Chair, International Conference on Spoken Language Processing, 1998
IEEE Spectrum magazine columnist and signal processing/neural network software reviewer
Selected Publications Related to Project


J. Bilmes, “Factored Sparse Inverse Covariance Matrices”. IEEE Int. Conf. on Acoustics Speech and Signal Proc., June 2000


Other Selected Publications


Author of over 30 scholarly works on Speech and pattern recognition, high-performance computing systems, parallel systems and languages, modeling human performance phenomena.

Synergistic Activities:

1. Associate Editor, IEEE Transactions on Multimedia
2. Currently developing a 4-course speech and language processing curriculum for UW (with M. Ostendorf), including graduate courses and a multi-disciplinary undergraduate course in speech interface design. Developed and am teaching a new modern course on graphical models in pattern recognition.
3. Member of the UW EE Computing and Planning committee.
4. Journalist and software reviewer for IEEE Spectrum
5. Designing and building a new UW EE speech-based supercomputer.

Advisors and Collaborators

Graduate Advisors: Nelson Morgan


Graduate Students & Postdoctoral Scholars Supervised in last year: Matt Richardson (PhD expected 02), Chia-Ping Chen (Ph.D. Expected '02), Gang Ji (Ph.D. Expeced 03), and Karim Filali (Ph.D. expected 04). A total of 4 graduate students supervised.
PROFESSIONAL PREPARATION

San Jose State University
Speech-Communication
BA, 1974
San Jose State University
Speech-Communication
MA, 1975
San Jose State University
Psychology
MA, 1976
UC Berkeley
Cognitive Psychology
PhD, 1983

POSTDOCTORAL INSTITUTIONS:
Max-Planck Institut fuer Psycholinguistik, The Netherlands

UC Berkeley, Institute of Cognitive Studies

APPOINTMENTS

January——–June 1998 Psychology Department, Holy Names College, Oakland, California, Lecturer: Statistical Methods (2 sections of 1 course)
1985———–present Institute of Cognitive Studies, UC Berkeley, Researcher
1983———–1987 Max-Planck Institut fuer Psycholinguistik, Visiting Scholar
January———–June 1979 Psychology Department, San Jose State University, Lecturer: Statistics and Experimental Design. (1 class)
June 1977———–May 1980 Psychology Department, UC Berkeley, Teaching Assistant (3 different courses) on Statistics and Experimental design.

PUBLICATIONS: 5 most closely related to the proposed project


**Publications:** 5 unrelated to the proposed project


**Synergistic Activities.**

I have written about principles of *discourse transcription*, and the theoretical assumptions underlying different methods. I worked on the markup standards proposed by the “Text Encoding Initiative” for spoken language encoding (Chapter P2.34 of the 1992 TEI final report). Markup standards enable the same transcript to be displayed in different ways for different research purposes.

I have written about *strategies when using computerized data archives* (Edwards, 1992) with an emphasis on strategies which can help avoid artefacts from the inevitable encoding errors which exist in probably all online corpora.

I compiled a multidisciplinary *survey of corpus-related resources* (Chapter 10, Edwards & Lampert,1993), and made it available via anonymous ftp. This survey can be found at various places on the web, and served (with permission) as a starting framework for Ken Litkowski’s ACL SIGLEX Resources page (http://www.clres.com/siglex.html).

**Collaborators.**
Dr. Collin Baker, Dr. Daniel Jurafsky, Dr. Willem J. M. Levelt, Dr. Peter R. Monge

**Graduate and Postdoctoral Advisors.**
First Masters Thesis: Dr. Peter R. Monge, Professor, Annenberg School for Communication, University of Southern California, Los Angeles, California
Second Masters Thesis: Dr. Robert Fox, Professor, Psychology, San Jose State University
PhD Dissertation: Dr. Seth Roberts, Associate Professor, Psychology, UC Berkeley

More Recent Research Advisors (within the past 2 years):
Dr. Susan Ervin-Tripp, Professor Emeritus, Psychology, UC Berkeley
Dr. Eleanor Rosch, Professor, Psychology, UC Berkeley.
NAME
Susan M. Ervin-Tripp

POSITION TITLE
Professor of Psychology Emeritus

EDUCATION
Vassar B.A 1949
University of Michigan M.A 1950 Social Psychology
University of Michigan Ph.D 1955 Social Psychology

HONORS
Phi Beta Kappa junior year
Margaret Floy Washburn Fellow 1949-50
Social Science Research Council Fellowship 1953-5
Guggenheim Fellowship 1974-5
Fellow of Center for Advanced Study in the Behavioral Sciences 1974-5
NAS China Delegation (Applied linguistics) 1977
US-France Scientific Exchange (European Science Foundation Project on Second Language in Migrant Workers) 1985-86
Cattell Fellowship 1985-86
University of California, Berkeley Faculty Research Lecture, 1994

RESEARCH AND/OR PROFESSIONAL EXPERIENCE
Research Assistant, Psycholinguistics Committee, SSRC, 1954-57.
Instructor, Harvard School of Education, 1955-58
Visiting Assistant Professor, Psychology Department University of California, Berkeley, 1958-59.
Assistant to Full Professor, Speech Department University of California Berkeley, 1959-74
Fellow, Center for Advanced Study in Behavioral Sciences, 1974-75
Professor, Psychology Department, University of California, Berkeley, 1975-1999
Professor Emeritus 1999-

CURRENT WORK
My work in the past ten years has focussed on two themes: First, the relation between contexts of language use and linguistic forms. The idea here is that language is learned from two sources: from the ability of learners to notice forms and their meanings, including contextual meanings, and to structure them into grammars, and from the pragmatic circumstances in which they hear and use forms, that is, what one might call the enabling conditions for language learning. Both are necessary for learning, but the relation between them has been left out of the picture. Second, the grammar of conversation, that is the structural organization of talk, including its development in monolinguals and bilinguals.

SELECTIVE BIBLIOGRAPHY
Closely Related Publications:


Significant Publications:


SYNERGY STATEMENT

A. Teaching:
Psychology 165  Language in social interaction. Included collection of new natural interaction text archive, and supervision of analysis of texts.

Psychology 128. Methods for the study of language in context. Included instruction in various types of transcription, coding, and conversational analysis.

Graduate seminars on transcription, using the archived tapes.
Graduate seminars on pragmatics, including speech acts such as requests (including mitigation according to social relations), narratives in conversation, humor in conversation


C. Organizations
President, International Pragmatics Association 2000-
HYNEK HERMANSKY

EDUCATION:

1972  Ing. (MS equivalent), Technical University Brno, Czech Republic,
1973-1978 Doctoral Graduate Studies, Technical University Brno, Czech Republic
1983  Doctor of Engineering, Electrical Engineering, University of Tokyo,

INTERESTS:

Recently, main efforts in studying and integrating human-like speech processing strategies into speech engineering systems. Among accomplishments belongs several widely used speech feature processing techniques which simulate basic properties of speech perception, such as Perceptual Linear Prediction (PLP) analysis, RelAtive SpecTrAl (RASTA) speech processing, or multi-band ASR.

ACADEMIC/PROFESSIONAL APPOINTMENTS:

1999- Senior Researcher, International Computer Science Institute, Berkeley, CA
1997- Professor, Electrical and Computer Engineering, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland, OR
1993-1997 Associate Professor, Oregon Graduate Institute
1992-1993 Senior Member, Technical Staff, U S WEST Advanced Technologies, Boulder, CO
1990- Fellow, International Computer Science Institute, Berkeley, CA
1988-1992 Member, Technical Staff, U S WEST Advanced Technologies, Boulder, CO,
1978-1983 Research Fellow, University of Tokyo, Tokyo, Japan
1975-1978 Assistant Professor, Technical University Brno, Czech Republic
1972-1975 Member of Research Staff, Technical University Brno, Czech Republic

PUBLICATIONS MOST RELEVANT TO THE PROPOSAL:


H. Hermansky and S. Sharma: "TRAPs - Classifiers of TempoRAI Patterns", Proceedings ICSLP'98, Sydney, Australia

H. Hermansky and M. Malayath: "Spectral basis functions from discriminant analysis", Proceedings ICSLP'98, Sydney, Australia


OTHER PUBLICATIONS:

H. Hermansky: "Should recognizers have ears?", Speech Communication 25, pp.3-27, August 1988
H. Hermansky and N. Morgan: "RASTA processing of speech, IEEE Transactions on Speech and
Audio Processing", Vol. 2, No. 4, pp. 578-598, October 1994


SYNERGISTIC ACTIVITIES:

Member of the Board, International Speech Communication Association
Associate Editor, IEEE Transactions on Speech and Audio Processing,
Member of Editorial Board of Speech Communication
Senior Member IEEE
Member, Acoustical Society of America

HONORS:


COLLABORATORS OVER PAST FIVE YEARS:

Dr. Nelson Morgan, ICSI Berkeley, California, Dr. Steven Greenberg, ICSI Berkeley, California, Dr. Ben Gold, MIT Lincoln Labs, MA, Prof. B. Yegnanarayana, IIT Madras, India, Prof. Herve Bourlard, IPFL Lausanne, Switzerland, Prof. Umesh Shrinivasan, IIT Kampur, India, Prof. L. Smekal, Technical University Brno, Czech Republic, Prof. Li Deng, University of Waterloo, Canada, Prof. Takayuki Arai, Sophia University, Tokyo, Japan (postdoctoral fellow), Prof. Noboru Kanedera, Ishikawa Institute of Technology, Ishikawa, Japan (postdoctoral fellow), Dr. Carlos Avendano, UC Davis, California (PhD. graduate), Prof. Hiroya Fujisaki, Tokyo Science University, Japan (PhD. adviser)
Kathleen R. McKeown

Professor and Chair
Department of Computer Science
Columbia University, New York, New York 10027
phone: (212) 939-7004
department@cs.columbia.edu
fax: (212) 666-0140
e-mail: kathy@cs.columbia.edu

Professional Preparation
A.B., Comparative Literature, Brown University, 1976.

Appointments
July 1997 to Present: Professor and Department Chair, Columbia University
July 1987 to July 1997: Associate Professor, Columbia University
September 1982 to July 1987: Assistant Professor, Columbia University
September 1980 to May 1982: Research Fellow, University of Pennsylvania
September 1978 to August 1980: IBM Research Fellow, University of Pennsylvania
September 1977 to May 1978: Teaching Fellow, University of Pennsylvania

Academic Honors
Fellow, American Association of Artificial Intelligence, elected April 1994.
NSF Faculty Award for Women, awarded September 1991 for a five year period.
IBM Faculty Development Award, August 1985, August 1986.

Five Most Closely Related Publications
Five Other Significant Publications


Synergistic Activities


Association for Computational Linguistics, President, 1992.

Association for Computational Linguistics, Vice President, 1991.

Co-chair, Program Committee, American Association of Artificial Intelligence, 1991.

Natural Language Area Chair, Program Committee, American Association of Artificial Intelligence, 1990.

Elected Officer, American Association of Artificial Intelligence, Spring 1987.

Chair, International Joint Conference on Artificial Intelligence (IJCAI) Natural Language Area Program Committee, 1987.

Graduate Advisor


Collaborators

Alfred Aho, Barry Allen, Shih-Fu Chang, Mukesh Dalal, Steven Feiner, Desmond Jordan, Judith Klavans, Karen Kukich, Johanna Moore, Rebecca Passonneau, Andrey Rzhetsky, Donia Scott.

Graduate Students

Regina Barzilay, Andrea Danyluk (Williams College), Michael Elhadad (Ben Gurion Univ.), Noemie Elhadad, Pascale Fung (HKUST), Vasilios Hatzivassiloglou (Columbia Univ.), Hongyan Jing, Min Yen Kan, Galina Datskovsky-Moerdler (MDY), Shimei Pan, Cecile Paris (CNRI, Australia), Dragomir Radev (IBM), Jacques Robin (Recife Univ.), Carl Sable, Barry Schiffman, James Shaw, Eric Siegel (startup), Frank Smadja (startup), Tony Weida (startup), Ursula Wolz (Trenton State) (total 20, 11 graduated).
LOKENDRA SHASTRI
http://www.icsi.berkeley.edu/~shastri

Education

Birla Institute of Technology and Science, Pilani, India.

Indian Institute of Technology, Madras, India.
Computer Science. MS. 1980.

University of Rochester, Rochester, NY.

Academic and Research Positions

1993–to-date Member AI Group, International Computer Science Institute, Berkeley, CA.

1985–1993 Assistant Professor and Research Scientist, Department of Computer and Information Science, University of Pennsylvania.

Five publications related to the proposed research


Five other significant publications


**Synergistic Activities**

- Workshop Organizer, “Connectionism meets AI” Penn/ARO, 1990; “Connectionist AI?” (with H. Gigley), IJCAI-89;
- Reviewer


**Funding agencies:** NSF, ARO.

**Research interests**

Artificial Intelligence, Cognitive Science and Computational Neuroscience: neurally motivated computational models of learning, memory, knowledge representation, and inference; spatio-temporal connectionist networks; the role of temporal synchrony in relational information processing; rapid memory formation in the hippocampal system; common-sense reasoning and its relation to abductive, deductive, probabilistic, and analogical forms of reasoning.

**Collaborators over the last 48 months**

Marvin Cohen (CTI), J.A. Feldman (UC, Berkeley), K. Grant (Walter Reed Army Medical Center), S. Greenberg (ICSI), J. Hobbs (SRI), R.T. Knight (UC, Berkeley), B. Miller (UCSF), S. Narayanan (SRI), D. Poeppel (U. Maryland), C. Schreiner (UC, San Francisco), and B. Thompson (CTI).

**Graduate Advisor**

J.A. Feldman (UC, Berkeley)

**Graduate and Post Graduate Advisees**


Elizabeth E. Shriberg

Speech Technology and Research Laboratory, SRI International
c/o International Computer Science Institute
1947 Center Street, Suite 600, Berkeley, CA 94704-1198
Phone: 510-666-2918, FAX: 510-666-2956
ees@speech.sri.com, www.speech.sri.com/people/ees/

Professional Preparation

1987 A.B., Linguistics                      Harvard College
1990 M.A., Cognitive Psychology          University of California at Berkeley
1994 Ph.D., Cognitive Psychology         University of California at Berkeley
1995 NSF-NATO Postdoctoral Fellow        Instituut voor Perceptie Onderzoek, The Netherlands

Appointments

Senior Research Psycholinguist, SRI International (1991–present)

Publications (available from www.speech.sri.com/people/ees/)


Additional Publications


   and across speakers: Predicting F0 targets when “speaking up”. *Proc. ICSLP*, Addendum Vol., pp.
   1–4, Philadelphia.

   California at Berkeley.


**Synergistic Activities**

1. **Current Editorships and Editorial Boards:**
   - Associate Editor, *Language and Speech* (Kingston Press)
   - Editorial Board, *Speech Communication* (Elsevier)

2. **Current and recent scientific committees:**


**Additional collaborators past five years**

Herbert Clark (Stanford U.), Robert Eklund (Telia and Stockholm U., Sweden), Horacio Franco (SRI), Larry Heck (Nuance Communications), Daniel Jurafsky (U. Colorado at Boulder), Nelson Morgan (ICSI), Patti Price (formerly SRI), Mari Ostendorf (Washington U.), Stefanie Shattuck-Hufnagel (MIT).

**Graduate and Postdoctoral Advisors**

Herbert Clark (Stanford U.), Susan Ervin-Tripp (U.C. Berkeley), John Ohala, (UC Berkeley), Patti Price (formerly SRI), Jacques Terken (IPO, The Netherlands).

**Thesis Advisor and Postgraduate-Scholar Sponsor**

Rebecca Bates (Boston University), Robert Eklund (Stockholm U., Sweden), Dilek Hakkani (Bilkent U. Turkey), Tim Paek (Stanford U.), Sergei Pakhomov (U. Minnesota), Madelaine Plauché (U.C. Berkeley), Gokhan Tür (Bilkent U., Turkey).
**Institutional Affiliations for all PIs and Senior Personnel**

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeff Bilmes</td>
<td>Dept. of Electrical Engineering, University of Washington</td>
</tr>
<tr>
<td>Jane Edwards</td>
<td>International Computer Science Institute</td>
</tr>
<tr>
<td>Dan Ellis</td>
<td>Columbia University</td>
</tr>
<tr>
<td>Susan M. Ervin-Tripp</td>
<td>University of California, Berkeley</td>
</tr>
<tr>
<td>Hynek Hermansky</td>
<td>Oregon Graduate Institute of Science and Technology</td>
</tr>
<tr>
<td></td>
<td>International Computer Science Institute</td>
</tr>
<tr>
<td>Katrin Kirchoff</td>
<td>Dept. of Electrical Engineering, University of Washington</td>
</tr>
<tr>
<td>Katherine McKeown</td>
<td>Columbia University</td>
</tr>
<tr>
<td>Nelson Morgan</td>
<td>International Computer Science Institute</td>
</tr>
<tr>
<td></td>
<td>Dept. of Electrical Engineering, University of California Berkeley</td>
</tr>
<tr>
<td>Mari Ostendorf</td>
<td>Dept. of Electrical Engineering, University of Washington</td>
</tr>
<tr>
<td>Lokendra Shastri</td>
<td>International Computer Science Institute</td>
</tr>
<tr>
<td>Andreas Stolcke</td>
<td>SRI International</td>
</tr>
<tr>
<td></td>
<td>International Computer Science Institute</td>
</tr>
<tr>
<td>Elizabeth Shriberg</td>
<td>SRI International</td>
</tr>
<tr>
<td></td>
<td>International Computer Science Institute</td>
</tr>
</tbody>
</table>