The Challenge of Spoken Language Systems: Research Directions for the Nineties

Ron Cole, Lynette Hirschman, Les Atlas, Mary Beckman, Alan Biermann, Marcia Bush, Mark Clements, Jordan Cohen, Oscar Garcia, Brian Hanson, Hynek Hermansky, Steve Levinson, Kathy McKeown, Nelson Morgan, David G. Novick, Mari Ostendorf, Sharon Oviatt, Patti Price, Harvey Silverman, Judy Spitz, Alex Waibel, Clifford Weinstein, Steve Zahorian, and Victor Zue

Abstract—A spoken language system combines speech recognition, natural language processing and human interface technology. It functions by recognizing the person's words, interpreting the sequence of words to obtain a meaning in terms of the application, and providing an appropriate response back to the user. Potential applications of spoken language systems range from simple tasks, such as retrieving information from an existing database (traffic reports, airline schedules), to interactive problem solving tasks involving complex planning and reasoning (travel planning, traffic routing), to support for multilingual interactions.

We examine eight key areas in which basic research is needed to produce spoken language systems: 1) robust speech recognition; 2) automatic training and adaptation; 3) spontaneous speech; 4) dialogue models; 5) natural language response generation; 6) speech synthesis and speech generation; 7) multilingual systems; and 8) interactive multimodal systems. In each area, we identify key research challenges, the infrastructure needed to support research, and the expected benefits.

We conclude by reviewing the need for multidisciplinary research, for development of shared corpora and related resources, for computational support and for rapid communication among researchers. The successful development of this technology will increase accessibility of computers to a wide range of users, will facilitate multinational communication and trade, and will create new research specialties and jobs in this rapidly expanding area.

R. Cole, H. Hermansky, D. G. Novick, and S. Oviatt are with the Oregon Graduate Institute of Science & Technology, Beaverton, OR 97006 USA.
L. Hirschman is with MITRE, Cambridge, MA 02139 USA.
L. Atlas is with the University of Washington, Seattle, WA 98195 USA.
M. Beckman is with the Ohio State University, Columbus, OH 43210 USA.
A. Biermann is with Duke University, Durham, NC 27706 USA.
M. Bush is with Xerox Palo Alto Research, Palo Alto, CA 94304 USA.
M. Clements is with the Georgia Institute of Technology, Atlanta, GA 30332 USA.
J. Cohen is with the IDA Center for Communications Research, Princeton, NJ 08540 USA.
O. Garcia is with the National Science Foundation, Washington, DC USA.
B. Hanson is with the Speech Technology Laboratory, Santa Barbara, CA 93105.
S. Levinson is with AT&T Bell Labs, Murray Hill, NJ 07974 USA.
K. McKeown is with Columbia University, New York, NY 10027 USA.
N. Morgan is with ICSI, University of California at Berkeley, Berkeley, CA.
M. Ostendorf is with Boston University, Boston, MA 02215 USA.
P. Price is with SRI International, Menlo Park, CA 94025 USA.
H. Silverman is with Brown University, Providence, RI 02912 USA.
J. Spitz is with the NYNEX AI Speech Group.
A. Waibel is with Carnegie-Mellon University, Pittsburgh, PA 15213 USA.
C. Weinstein is with the Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA 02173, USA.
S. Zahorian is with Old Dominion University, Norfolk, VA 23529 USA.
V. Zue is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.
IEEE Log Number 9506773.

I. INTRODUCTION

A SPOKEN language system combines speech recognition, natural language processing and human interface technology. It functions by recognizing the person's words, interpreting the sequence of words to obtain a meaning in terms of the application, and providing an appropriate response back to the user. Potential applications of spoken language systems range from simple tasks, such as retrieving information from an existing database (traffic reports, airline schedules), to interactive problem solving tasks involving complex planning and reasoning (travel planning, traffic routing), to support for multilingual and multimedia interactions.

Spoken language systems make it possible for people to interact with computers using speech, the most natural and widely-distributed human mode of communication. Although these systems are still in their infancy, they have the potential to revolutionize the way that people interact with machines. Because spoken language systems will support human-machine interaction in a natural way that requires no special training, these interfaces will eventually make computer-based resources available to many new groups of users (casual users, telephone users, hands-busy or eyes-busy users, handicapped users, users with a different native language), as well as supporting expert users in handling information-intensive problems.

Spoken language systems technology has made rapid advances in the past decade, supported by progress in the underlying speech and language technologies as well as rapid advances in computing technology. As a result, there are now several research prototype spoken language systems that support limited interaction in domains such as travel planning, urban exploration, and office management. These systems operate in near real-time, accepting spontaneous, continuous speech from speakers with no prior enrollment; they have vocabularies of 1000–2000 words, and an overall correct understanding rate of almost 90% [96], [5], [110], [4], [26].

Although progress over the past decade has been impressive, there are significant obstacles to be overcome before spoken language systems can reach their full potential. Systems must be robust at all levels, so that they handle background or channel noise, the occurrence of unfamiliar words, new accents, new users, or unanticipated inputs. They must exhibit more "intelligence," knowing when they don't understand or only partially understand something, and interacting with the user appropriately to provide conversational repairs and graceful degradation. They must integrate speech with other modalities,

deriving the user's intent by combining speech with facial expressions, eye movements, gestures, handwriting, and other input features, and communicating back to the user through multimedia responses. Finally, to reach their full potential, they must be multilingual, performing speech-to-application translation or even speech-to-speech translation.

What research is required to produce such spoken language systems, and what infrastructure i s required to support this research? These questions were considered by a group of scientists at a two-day workshop in February, 1992, sponsored by the National Science Foundation. This article is based on a report written by the workshop participants.

The article focuses on fundamental problems in spoken language understanding—key research areas in which progress must be made to produce systems that are accurate, robust, and graceful. We should keep in mind, however, that spoken language technology is a vital and expanding industry, with demonstrated success in a growing number of application areas. Moreover, systems need not be "perfect" to be successful, and incremental improvements in system performance are leading to increasing acceptance of the technology. Research that improves the capabilities of current systems is important to progress in the field, and is certain to contribute to advances in future systems.

In the following section, we identify eight areas in which fundamental research is needed. In each area, we discuss the nature of the problem and the key research challenges. In Section III, we consider the infrastructure that is needed to support the research. In the final section, we consider some of the anticipated benefits of spoken language systems.

II. RESEARCH DIRECTIONS

We identify eight areas in which fundamental research is needed to produce spoken language systems. These are:

- **Robust Speech Recognition**—to provide graceful degradation when the system loses information due to limited bandwidth, background noise, channel distortion, etc.
- **Automatic Training and Adaptation**—to make systems easy and cheap to adapt or train in new domains.
- **Spontaneous Speech**—to model the prosody of spontaneous speech, pauses, hesitations, repairs, and turn-taking behavior.
- **Dialogue Models**—to enable spoken language systems to carry on a coherent conversation with the user.
- **Natural Language Response Generation**—to provide coherent, appropriate output to the user.
- **Speech Synthesis and Speech Generation**—to produce comprehensible speech output to the user and to enhance our understanding of speech.
- **Multilingual Systems**—to provide multilingual information access and speech-to-speech translation.
- **Multimodal Systems**—to increase the accuracy and naturalness of human computer interaction by integrating speech with other sources of information, such as facial expressions, gestures, and handwriting.

A. Robust Speech Recognition

Robustness in speech recognition can be defined as minimal, graceful degradation in performance due to changes in input conditions caused by different microphones, room acoustics, background or channel noise, different speakers, or other small (insofar as human listeners are concerned) systematic changes in the acoustic signal.

At present, speech recognition systems are not very robust. Their performance degrades suddenly and significantly with modifications as minor as a change in microphone or telecommunication channel [48]. Systems trained in the laboratory fail when exposed to operating conditions in the field [146]. Users will naturally be reluctant to rely on automatic speech recognition if they have to talk in a highly constrained way, if it fails on a day when they have a cold, or if performance drops severely when there is a reasonable level of background noise.

Although signal processing strategies show promise in leading to robust systems [101], [62], [45], [88], [79], [131], [13], [89], [21], [1], [35], [132], [47], [52], [20], the fundamental method for improving robustness is to understand better the many sources of variability in the speech signal. Figure 1 shows some of the many sources of variability in the speech signal from the viewpoint of a machine recognizer. Variability is typically due to the talker and the nature of the task, the physical environment, and the communication channel between user and the machine.

Some of these sources of variability are clearly irrelevant to the task and should be treated as noise. Other sources of variability, such as speaking rate and fundamental frequency, provide information which contribute to the meaning of an utterance; these should be treated as knowledge sources rather than noise sources. In each case, there is a continuing need for new ideas and innovative approaches leading to algorithms which are sensitive to the variabilities of interest and remove or discount irrelevant sources of variability.

Key Research Challenges and Fundamental Scientific Issues:

1) Modeling Coarticulation and Phonetic Context: The spectral characteristics of a sound segment vary tremendously from one linguistic context to another (see [105] for a partial overview). At present, coarticulatory variability due to segmental context is mostly accounted for by context-dependent hidden Markov models; this vastly increases the number of classes on which a recognizer must be trained.

More explicit models of coarticulation could improve this situation. Very often, coarticulatory variability is conditioned by prosodic structure and arises through systematic modifications in timing and rates of movement of the articulators. Deeper understanding of the articulatory dynamics of the speech production process and of its influence on the resulting speech signal could help here.

The acoustics of the speech signal depends on a variety of factors besides the intended sequence of phonemes to be communicated—factors such as the geometry of the particular speaker's articulators, dynamic and mechanical constraints on their motion, various prosodic effects related to intention, etc.

In the simplest cases (e.g., isolated-word or speaker-dependent recognition), it is possible to ignore these "extraneous variables" completely, and in more complicated scenarios much success has been obtained by taking them into account only implicitly, using statistical modeling techniques.

As we attempt to improve the power and robustness of speech-recognition algorithms it may be necessary, however, to take into account the underlying variables of the speech production process in a more explicit fashion. Two classes of approaches have been developed along these lines. On the one hand, researchers have modeled speech production as a generic dynamic system [60]. Standard tools from system identification are then used in an attempt to derive a mathematical description of this process. On the other hand, the known properties of human articulatory dynamics can be used to constrain the mathematical description of the system dynamics [111], [107], [17]. In both classes of systems one attempts to achieve more robust speech recognition by first deducing a certain amount of information regarding the dynamics underlying speech production.

2) Modeling Speech Rate: The effects of tempo are poorly understood. What factors cause speakers to speak more quickly or more slowly? If the effects of tempo on the production process were better understood (see e.g., [33], [93], [133]), local changes in speaking rate might be used to recognize such prosodic patterns as stress or phrase-final lengthening (e.g., [122], [19], [103]); more global changes might help parse changes in topic or conversational turn (see e.g., [8], [7], [55]) and even some more intricate pragmatic differences among utterances ([56]).

3) Modeling Speaker Differences: Speaker differences arise from two different sources: a) differences in anatomy of speech production organs; and b) differences in acquired speech production skills.

Until quite recently, modeling of speaker-dependencies was largely ignored and speaker-independent recognition was accomplished by training on large amounts of data from many different talkers. Recently, some systems have begun to model specific talker characteristics in a more explicit way, in the context of research on speaker adaptation [44], [141], [108].

Progress in modeling speaker differences can benefit from basic research on talker characteristics arising from anatomical differences from fields as diverse as speech physiology, phonetics, psychoacoustics, and speech synthesis (e.g., [12], [81], [129], [32], [91], [43], [130], [87], [90], [102], [61], [65], [64], etc.). Incorporating these results into recognition systems could lead to new models and representations, with implications not only for speech recognition but also for talker verification and identification. Recent works indicate that even simple engineering models of human speech perception [46], [49] could alleviate some of the dependent variability.

Additional research is also required to understand dialect differences and other effects of social context on speech production (e.g., [69], [42], [41], [51], [70], [59], [11], [18]). Incorporating explicit representations of phonological differences among dialects could vastly reduce the amount of training data required for training of a robust recognizer. Explicit modeling of such dialect variation at the phonetic level will be particularly important if it turns out, as suggested by [92], that patterns of coarticulatory variation across word boundaries can differ from one dialect to another.

4) Acoustic Environment and Microphone: It is widely accepted that users do not want to be encumbered by a headmounted or handheld microphone when communicating with a machine. Thus, it is highly desirable to have a remote microphone attached to the system, which can track a talker and maintain consistently high signal quality.

Multiple microphone systems offer the prospect of being able to track a remote talker [30], [114], [15], [6], [31], [115], [116]. This approach has yielded some success, but considerable work is still necessary to handle speech acquisition in a free space. Some of the hardest problems arise from reverberation. Algorithmic solutions require some form of deconvolution, which is a very difficult procedure; therefore, spatial filtering and mechanical/acoustic augmentations will probably be required. Finally, in working environments, there may be "interference" from the speech of other talkers. This is a most difficult problem due to the spectral similarity of the interference.

Many current recognizers require a specific microphone for good performance. Research is now underway to understand how to adapt a system trained on one microphone to perform at full capability using a different microphone/environment [48], [52], [119].

5) Communication Channel: Speech recognition over telephone channels is imperative. The demand for this technology is increasing dramatically, due to economic pressures, the successful deployment of systems that save hundreds of millions of dollars per year in operating costs, and the rise of mobile cellular telephony, which provides a natural domain for many commercial applications of human-computer interface technology.

When communicating over current telephone channels, many previously ignored phenomena such as echoes, noise, nonlinearities, and spectral distortions arise and need to be addressed. Recent work indicates that at least partial alleviation of some of these effects during speech feature extraction is possible (see, e.g., [50]).

6) Models of human speech perception: It is reasonable to assume that the properties of human auditory perception have influenced the coding of linguistic information in the speech signal. That is, one would expect that speech components enhanced in human hearing would be the components primarily used in decoding the linguistic message in speech. However, many properties of human auditory perception are not well represented by the short-term spectral analysis in the front end of a typical automatic speech recognizer.

In order to resolve many of the inherent limitations of current recognizers, better fundamental representations of the signal must be formulated. Such transformations, or feature extraction methods, will mitigate many of the problems arising from the sources of variation depicted in Fig. 1 (i.e., talker, environment, channel, etc.). The development of auditory models for speech processing is still at an early stage, but current work in the application of some of these models to automated speech recognition appears quite promising.

Some recently developed speech analysis techniques attempt to model the basic properties of human speech perception [109], [34], [75], [46], [50]. These techniques can provide significant improvement of recognition robustness by alleviating some of the nonlinguistic sources of variability in speech, such as differences due to talkers, differences in the acoustic environment and in background noise, or overall spectral differences due to a change of microphone or microphone position [48], [52], [13].

Models of speech perception for processing stages beyond the periphery have been proposed, but little attempt has been made to incorporate concepts of these models into systems for speech recognition (e.g., [72], [126], [120]). Understanding human speech perception is an important step in the development of spoken language systems.

7) Confidence and Rejection: In real applications, a speech recognition system must deal with unexpected or unusual input. The speaker may produce words that are not in the recognition vocabulary, or simply pause after a system prompt, in which case the system may be presented with background sounds from a radio or television set. To be useful in real world applications, speech recognizers must gauge the confidence of words that are recognized. Without measures of confidence, spoken language systems produce unacceptable errors, and are unable to engage the speaker in graceful dialogues.

Measuring confidence is an unsolved problem. Current speech recognizers do not know what they do not know. They produce unreasonable responses, such as mistaking background noise for speech, or recognizing words that have an entirely different prosodic structure from the words spoken. Simple approaches, such as applying rejection thresholds to recognition scores (based on training data) are found to be highly sensitive to background noise, and vary widely from speaker to speaker. Basic research is needed to develop robust measures of confidence that use all available information.

B. Automatic Training and Adaptation

The introduction of spoken language systems into a variety of real world applications requires fundamental research on how to adapt these systems quickly and cost effectively to new applications. Currently, the high cost of porting such systems to new applications represents a major obstacle to widespread deployment. For speech recognition, highly effective automated training procedures have been developed, but these require large amounts of task-specific data for reasonable performance. For example, in the ARPA Air Travel (ATIS) domain, joint data collection activity across five sites resulted in the collection of over 14,000 utterances [23]. The data collection activity represents several person-months of effort at each site to collect and transcribe the data, not to mention the costs of checking and distributing the data.

Natural language understanding systems require training data too, but they also rely heavily on computational linguists to build the lexicon, to tune and debug the grammar, to provide the domain model, and to link the domain semantic rules to the domain model. This process is not only labor intensive but also requires scarce expertise.

In addition, we have no set of metrics for portability. The current evaluation procedure for language understanding in a single domain is expensive—it requires careful definition of terms (e.g., what does "evening" mean in the travel domain), followed by manual specification of a "correct answer" for each utterance. Current annotation proceeds at the rate of about 100 utterances a week for a trained annotator, and any evaluation requires thousands of annotated training utterances, in addition to test data.

Multilingual systems represent a significant challenge for portability as well. Porting a system to a new language often places an additional burden of "language independence" on all components of the system [36], in addition to the need for new training data for both speech and language, although there are architectures that can make the port less difficult than it would otherwise be (see, e.g., [25]).

Until we develop faster, less labor intensive methods of porting or adapting systems to new domains and new languages, the applicability of spoken language systems will be restricted to a very small, carefully chosen set of high-return applications; it is simply too expensive to proliferate applications.

Key Research Challenges and Fundamental Scientific Issues:

1) Better Use of Training Data: One approach to the problem of collecting new training data for each application is task-independent vocabulary modeling [58]. This approach could be extended to task-independent language modeling, and rapid adaptation to new domains using task-independent data supplemented by only a small sample of task-specific data, as suggested by recent work in cache-based language modeling [68].

2) The "New Word" Problem: The occurrence of unknown or out-of-vocabulary words is one of the major problems frustrating the use of automatic speech understanding systems in real world tasks. Real users of spoken language systems cannot be expected to know exactly what words are in the system lexicon, and will often produce words that are unknown to the system.

To detect new words in the input is one of the most difficult steps in the process. For example, it is not sufficient to determine that an area of the input is poorly matched; it is necessary to differentiate a new word from background speech of other talkers, from breath noises, coughs, filled pauses, and from environmental noises such as telephone rings and door slams. Once an unknown word has been identified, it must be added to the recognition vocabulary, which involves generating a spelling for the word automatically (if printed text is required), determining its pronunciation, and constructing a word model. In order to be included in future searches, the word must also be added to the system's language model. This usually means determining the class membership of the word, since most recognition systems use some form of word classes in their language models.

3) Discovery Procedures for Syntactic and Semantic Classes: Natural language systems typically require several kinds of classification for words. If parsing is involved, words need to be marked for part of speech (and other syntac-

Talker

Talker characteristics
Social context/task
Dialect
Phonetic Context
Prosodic Context
Rate
Mood/Style

Acoustic Environment

Other Talkers
Background Noise
Reverberations

Microphone

Non-linearities
Spectral Characteristics
Electrical Noise
Directional Characteristics

Channel

Echoes
Dropouts
Added Noise
Spectral Distortion

Capture

Acoustic Processing

Machine Listener
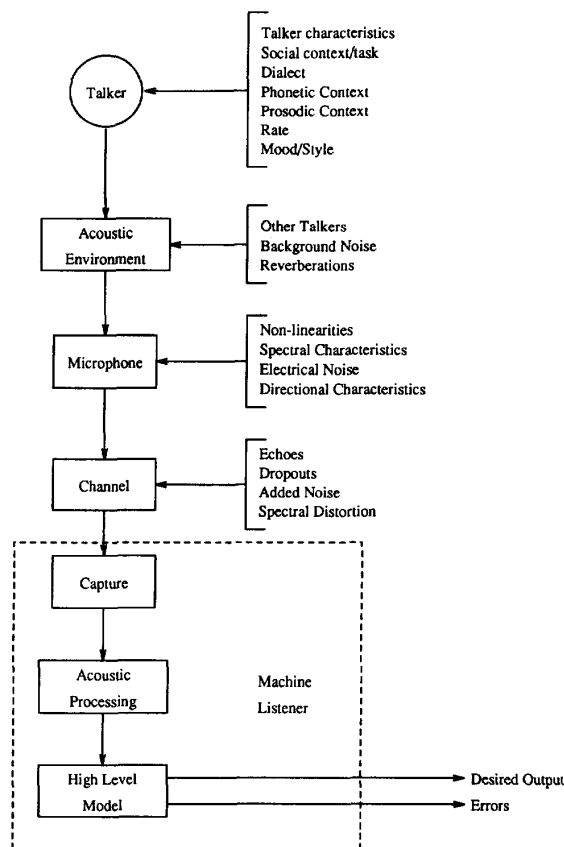
High Level Model

→ Desired Output
→ Errors

Fig. 1. Sources of variability in the speech signal.

tic information, such as complement structure). In addition, semantic information is needed, as well as their mapping into the "back-end", e.g., we need to know that "Philly" maps into "Philadelphia" for purposes of accessing air travel information. Since some portion of the vocabulary tends to be quite application specific, there is a need to automate as much of this as possible. Automatic part-of-speech tagging and automated discovery of syntactic and semantic classes will aid in porting to new tasks and also between languages.

4) Knowledge Engineering Bottleneck: One of the most labor- and expertise-intensive tasks is the construction of a domain model, which provides semantics for the objects in the domain, along with a taxonomy and a specification of the objects' relations to each other.

To decrease the cost of portability, we must find ways to utilize existing repositories of "expert information" such as thesauri and lexicons, and semantic representations such as WordNet [86]. Under the Consortium for Lexical Research[1] and the Linguistic Data Consortium,[2] these resources are being made widely available, but we need further research on how to extract and utilize the information contained in them.

[1] E-mail address: lexical@nmsu.edu.

[2] E-mail address: ldc@unagi.cis.upenn.edu.

Another serious problem is the linkage of the domain model to the application back-end (e.g., a database using SQL input) and to the lexicon or the lexical semantics. There has been relatively little research on automation of this process, particularly in the context of building a spoken language system.

5) Graceful Degradation and Knowing What You Don't Know: There is always a trade-off between depth of modeling and robustness—shallow models are easier to build but also provide more limited understanding. If it were possible to model a system's boundaries better, that is, what it doesn't know, in addition to what it does know, it might be possible to get by with shallower models, but also to provide better feedback to the user and more graceful error degradation.

6) Evaluation of Portability: In order to measure progress in portability, it is important to find some reasonable metrics that are themselves fairly cheap to implement. This may require new ways of doing system evaluation, since the current evaluation methods measure "understanding" and are expensive to implement for a single domain (cf. MUC [123] and ATIS [23]), let alone for multiple domains. Until we find reasonable and affordable metrics of portability, we will see little progress in this difficult area.

C. Spontaneous Speech

Benefits of Spontaneous Speech: The ability to deal with spontaneous speech phenomena is an important property of robust systems. Systems that cannot be used in a natural manner will not find general acceptance. Research in spontaneous speech will allow computers to repair conversational breakdowns and misunderstandings and will liberate users of spoken language systems from static, stilted interfaces by enabling more natural dialogue interaction. Spontaneous speech is notoriously problematic, full of "improper" usages, mismatched agreements, run-on sentences, and hesitations and restarts which interrupt words and grammatical constructions. Such interruptions and inconsistencies go mostly unnoticed by the participants in conversation. Many people are quite surprised to see a literal transcription of what they have said. The conversants handle their interchanges effortlessly, in the way they take turns, make interruptions, detect and correct misunderstandings, and resolve ambiguous references. How can these processes of control be modeled formally in a manner sufficient to bring this sort of coherence to computer understanding of spontaneous language? To what extent are these capabilities needed to build successful human-machine interfaces?

Socio-linguistic research in conversational analysis has described a wide range of conversational characteristics which are not directly representable in sentence-level and other text-oriented accounts of conversation. These characteristic behaviors include: 1) lower level events such as pauses, filled pauses (e.g., "uh"), laughter and other nonspeech noises (inhalation, cough); 2) suprasegmental phenomena, such as speaking rate, pitch, and amplitude; 3) meta-sentential events such as correction and editing ("Denver, I mean BOSTON"); 4) back-channel communication that is critical in communicating that the hearer is present and/or paying attention; and

5) nonverbal communication (eye contact, nodding) that plays a part in maintaining a conversation.

Because prior research focused on read speech and written text, our knowledge of spontaneous speech is still limited. We know that there are regularities associated with spontaneous speech phenomena that do not appear as frequently in read speech. For example, repairs occur with some frequency (around 6% percent of sentences in rather planned spontaneous data such as ATIS [113], and in 34% of sentences in a human-human dialogue corpus [71]), but occur more rarely in read material. Such repairs are easily recognized by humans, but our current spoken language models are not rich enough to handle them.

Prosody is another important component of spoken language that is not well represented in written language. Understanding speech without prosody is like understanding written text with no punctuation. Without prosody, we lose the cues that make spoken language coherent in spite of the high rate of disfluencies and ill-formed constructs. In fact, prosody may enable spoken language to convey more information than written text. In human-machine interactions, prosodic cues may provide valuable information for computational models with limited semantic knowledge, even though the cues may be only redundant information for human listeners with a detailed knowledge of the world. In addition, prosody is a limiting factor in speech synthesis applications [66].

There are three major reasons for the problems of current models. First, the handcrafting of language understanding systems leads to a *competence-based* model rather than a *performance-based* model. Second, we do not understand well how to treat various phenomena as information rather than noise. This is particularly true of things like change in speaking rate, or hesitations in speech. Third, because we do not capture the information contained in these phenomena, we do not know how to normalize the utterances, e.g., how to use hesitation to locate a repair and correct for it. By studying these phenomena and by explicitly incorporating them into both the acoustic processing (for detection) and the language processing (for interpretation), we should be able to build much more robust spoken language systems.

Many of these behaviors contribute to monitoring and regulating the conversation itself. For example, the listener may provide periodic back-channel acknowledgement of understanding, the speaker can signal completion of a query or comment, the listener may wish to signal a desire to take the floor, or to request a clarification, etc. The ability to carry on a multiparty conversation depends critically on exchanging these signals: turn-taking, correction of errors, request for clarification, and listener confirmation are all necessary for successful communication. Research on turn-taking dynamics and conversational control will contribute to our understanding of human-human interaction, and will also make important contributions to building better, more natural and usable human-machine interfaces.

Key Research Challenges and Fundamental Scientific Issues: Spontaneous speech phenomena are not simply "linguistic chaff" to be discarded, but kernels of linguistic action that have meanings and purposes that are helpful—maybe even necessary—in understanding spoken language. The fundamental scientific issues in spontaneous language involve development and testing of theories of linguistic interaction that account for the observed behaviors. In particular, such theories need to be expressed in computational terms, so that spoken language understanding systems can better extract useful information from the range of linguistic and extra- and meta-linguistic phenomena associated with spontaneous speech. These theories of interaction may well have significant implications for the design and development of human-computer interfaces. That is, understanding spontaneous speech phenomena may help to elucidate the underlying principles of communicative interaction. Accordingly, research in spontaneous speech should stress the following issues:

1) **Computational Models of Spontaneous Speech:** There has been relatively little work on spontaneous speech, in particular, on computational models of spontaneous speech. Such models are necessary at all levels—acoustic, linguistic and prosodic—in order to detect and account for these phenomena in automatic speech understanding. Further, we need to understand the conditioning factors that increase or decrease their appearance, so that we will know how to either model them or minimize them in appropriate interfaces. Finally, such models need to be integrated into architectures for spoken language processing, in both speech understanding and generation components.

2) **Prosodics:** We know that prosody can provide information that helps humans understand speech (e.g., in read speech by radio announcers [103]). We also know that prosody of read speech differs from spontaneous speech [16]. Recent results are beginning to show that use of prosody can aid automated understanding of spontaneous speech [134]—provided that the system can detect prosodic phenomena reliably [139], [140] and can correlate these prosodic cues with higher-level syntactic, semantic, pragmatic, and conversational structures. For example, phrasal prominence may help to detect high-information regions of a discourse, for use in automatic gisting and summarization [10]. However, this is a very young research area and more research is needed to understand how people use prosodic information, and to understand how prosodic information can improve the performance of spoken language systems, both for recognition and for generation (discussed below).

3) **Understanding Conversational Dynamics:** We need to know how turn-taking models can account for coordinated speech and simultaneous speech, and how turn-taking might be signaled acoustically, e.g., via pauses, lengthenings and pitch patterns. We need to know how spontaneous speech phenomena can be used by a speaker and by a listener to deal with interruptions and seeming "irregularities" in utterances, or to signal certain kinds of conversational interaction. We also need to understand the effects of differences in modality of communication on conversational control acts and the factors that determine the limits of acceptable ambiguity and uncertainty in conversation.

4) **Evaluation:** The research community needs adequate metrics to evaluate how well systems can handle various spontaneous speech phenomena and issues of turn-taking in

conversation. To do this, we need appropriately annotated corpora and representative test suites to evaluate the importance of these phenomena and to track our progress in accounting for them. Specifically, this will require spontaneous speech corpora with detailed transcriptions (including prosodic annotation), at least some of it collected in "two-party" conversation settings (like the SWITCHBOARD corpus being collected at Texas Instruments [37]). Efforts are currently underway to create a standard notation for prosody, so that training and evaluation materials can be prepared [117], but this is only a first step towards a corpus of prosodically labelled spontaneous speech.

D. Dialogue Models

Dialogue processing is the enabling technology for spoken language systems. While speech recognition technology may provide better and better hypotheses about what words were uttered, the machines will not properly use these hypotheses unless they extract the user's meaning from those tokens and efficiently respond to the user's needs.

True speech understanding requires that individual utterance meanings be understood in the context of the larger dialogue structure [2], [40]. This structure must coordinate a variety of information, including the ultimate goals of the interaction, the subgoals being attempted, the status of the system knowledge base, models of user knowledge, and a history of the interaction. A full specification of the utterance meaning includes its connection to the overall task structure as well as its relation to the system and user model knowledge bases.

Information involved in utterance understanding flows in two directions. The result of the understanding process is a kind of unification of knowledge fragments from the utterance and from the system knowledge bases. Individual utterances supply specific pieces of knowledge about the state of the world (in the context of the specific task domain). These must be connected to global data structures (from the system knowledge bases) which describe how entities are related to each other in the domain. These global data structures are used to complete the individual sentence meaning. An example of the unification process occurs in the exchange:

COMPUTER: What is the switch setting?
USER: It is up.

The machine's output could be represented by *state(switch, Y)* and the user's response by *state(X, up)*. The total meaning of the user's response is *state(switch, up)*, which is obtained by integrating (unifying) information from the utterances of the computer and user, based on discourse level information about question/answer pairs and reference.

In general, large amounts of information at the dialogue level need to be accessible to understand the meaning of the utterance in context. Thus, the resolution of noun phrases (especially pronouns), the processing of elliptical constructions, the selection of appropriate meanings for verbs and scale words, and many other sentence level structures can only be handled by properly finding linkages to higher level dialogue structures.

By understanding and using dialogue constraint, it will be possible to build more robust and more user-friendly systems.

This will happen in several ways. First, dialogue modeling can provide improved error correction for the recognizer. The dialogue system can provide expectations for the incoming utterances that will improve recognition rates. Second, it can provide improved total system robustness. When major or minor errors occur in an interaction, the dialogue system will persistently seek achievement of the goal. Third, it can provide improved system efficiency. The use of a domain model, dialogue structure, variable initiative, intelligent error handling, and user modeling all contribute to reducing the amount of user input needed to do the job and increasing the rate at which the interaction will converge on the goal. Given the proper dialogue model, the user need only provide short fragmentary utterances to guide the system through the appropriate subdialogues. System outputs will avoid repeating knowledge known to the user and deliver only essential information needed for effective forward movement. These benefits are not second-order in effect; they are dramatic in their influence on total system behavior and are required for spoken language systems to come into common use, because human users expect systems to participate in cooperative dialogue.

Key Research Challenges and Fundamental Scientific Issues:

1) Discovering the Structure of Dialogue: Typical dialogues are usually organized into a series of subdialogues, each of which is aimed at solving a particular subgoal [39], [73], [104]. The individual subdialogues provide what is called "focus" [39], and the tracking of subdialogues is called "plan recognition" [3]. The relationships between the subdialogues are often quite complex, some being nested within others, some being functionally disjoint from others, and so forth. This nesting affects not only content and referential structure, but prosodic structure as well [54]. In order to understand and participate in conversational interaction, the dialogue/subdialogue structure must be correctly understood and modeled.

2) Using Dialogue Structure in Speech Recognition: The dialogue model provides, at each instant of time, a powerful expectation of what is to be said next. The currently active subgoal will make very strong predictions, and other locally nonactive subgoals will make weaker predictions. The combination of all the information from the dialogue level can substantially sharpen estimates for improved recognition [22], [136], [138].

This leads to a new formulation of the speech understanding problem. Instead of receiving an acoustic input and passing a meaning to the higher level, the recognizer could receive both the acoustic input and a representation of expected meanings. The output of the recognizer should be a best guess of which of the expected meanings was, in fact, received. This model of speech understanding could reduce perplexity and provide improved error correction [143].

3) Building a Variable Initiative Capability into the Processor: The possibility of moving from subdialogue to subdialogue in nearly arbitrary ways leads to the question of who controls these transitions [118]. The answer is that an efficient dialogue capability requires that either participant be able to take control. If one participant, machine or human, has most of the knowledge related to a subtopic, efficiency

may require that that entity dictate dialogue transitions to properly guide the interaction to success. However, in typical cooperations, each participant will have dominant knowledge on particular subtopics, so control needs to be passed back and forth. Thus a machine needs to be able to function in "passive mode," which obediently tracks the preferences of the user, or "directive mode," which insists on leading the user through its own agenda. Intermediate levels are also useful where the machine may yield control to the user while injecting suggestions along the way or where the machine may gently take control while respecting user preference.

A system that allows several levels of control is said to demonstrate "variable" or "mixed" initiative. One can expect variable initiative to be superior to fixed initiative in typical problem solving. An example situation where variable initiative is important occurs in the case where a novice needs, at first, to be pedantically led through a series of steps (machine directive mode) but later can take initiative (machine passive mode) on a growing set of subtasks as he or she learns to function in the environment.

4) Incorporating a Model of the User: A key aspect of a dialogue system is its model of the user [29], [67], [97]. Processes of input recognition, output generation, and internal decision-making all depend on user modelling. Word usage, grammatical constructions, and transmitted meanings will differ for users of different backgrounds and different levels of expertise. A user model must contain both stable long-term information and a fast-changing short-term record of the current interaction. The long-term information relates to the vocabulary and abilities of the user; the short-term information tells what the user has learned in the immediate past so that the machine can continually account for it. An example of long-term information is the assertion that a user knows how to measure a voltage; an illustration of short-term information is the case where a user has just been told where a particular object is.

5) Error Handling: A critical part of dialogue-based interaction is the ability of the participants to ask questions and clarify responses, so that they interactively refine their understanding until a point of mutual intelligibility is reached. Spoken language systems will be expected to provide such capabilities, especially as they become more sophisticated. There are many open questions concerning spoken language systems and error handling; for example, what is the best way to handle a partially understood sentence? Should the system guess, should it report what it understood, should it ask the user to repeat or rephrase the question? When the system does make a mistake, how should the system present its response, to help the user diagnose a possible system misunderstanding? What is the cost of an error [57]? Graceful error handling, clarification dialogue and detection and correction of presupposition failures are critical features for a spoken language system.

6) Generation of Appropriate Output: Another important part of a dialogue system is its output generation facility [78], [76], [84]. This may be in a typed, voiced, or graphic mode, and its purpose is to enunciate the machine's portion of the interaction as dictated by the dialogue processor. Efficient output will code the meaning of the message to be transmitted in a manner that properly accounts for the user's knowledge. Generation of appropriate output is discussed further in the next two subsections, on response generation and speech synthesis.

E. Natural Language Response Generation

A spoken language interface involves more than just recognition and interpretation. An interface must engage in two-way dialogue between user and system. Interpretation alone does not allow the system to respond to the user in an intelligible and helpful way. Research into response generation aims at determining the content and form of the response so that it is actually useful. A response that contains far more information than is needed requires a user to expend additional energy sifting through information for the piece of interest [142]. Conversely, a response containing too little information can mislead or derail a user in the problem solving process.

Although response generation is a critical component of interactive spoken language systems, and of any human computer interface, very little research in these areas is currently funded in the United States. Instead, current funding efforts assume that once a spoken utterance is interpreted, the response can be made using the underlying system application (e.g., the results of a database search) and commercial speech synthesizers. These efforts ignore the results of natural language research in the early 1980's which showed why such an approach is inadequate [63], [53], [80], [83], [85].

In any interactive situation, a system must be able to interpret input and take some action that achieves what the speaker intended. Without a response generation component, this must be an action that the underlying back-end application system can carry out. Previous work has shown, however, that for a variety of different applications this is an unrealistic expectation. For example, in an interface to a database system, such response would be limited to results of a search of the database. But there are many types of requests that cannot be handled by searches or other underlying system capabilities. For example, it has been shown that users would like to ask questions about the type of information available in the underlying database, or questions requesting the definition of terms, or questions about the differences between concepts [77], [128]. These questions cannot be answered unless the system includes facilities to determine what information to include. Given that this information does not directly mirror the user's question, the system also needs to determine how to phrase the information in language. Similarly, expert system explanation is another application where it has been shown [127] that a simple "translation" of the underlying inference trace (as is often done using templates [112]) does not produce a satisfactory explanation of the system's reasoning. Finally, in machine translation, where the content of the generated text is determined by parsing the source language, generation techniques are required to select the wording that correctly conveys the original meaning.

Key Research Challenges and Fundamental Scientific Issues: Research in language generation spans a variety of issues. It addresses the problem of what information should be

included in a response as well as how the information should be organized. For example, the system needs to determine which information should come first and how internal pieces are related to each other (e.g., coordinated or subordinated). Language generation also requires determining the form of the response, including the words and the syntactic structure, or ordering of the words in a sentence. Each of the research challenges below impacts on all of these generation tasks:

1) **Generation as Part of Dialogue:** When generation takes place as part of an interactive dialogue system, responses must be sensitive to what has already been said in the current session and to the individual user. The past history influences the content of the response; the system should avoid repetition and provide information that is relevant to the user's goals and background knowledge. It influences the form of the response, since the system needs to select vocabulary that the user can understand. Furthermore, knowledge about what information is new, or not previously mentioned, and what information is given, or available from previous discourse, can influence word ordering. While there has been some work addressing these issues, the influence of discourse on response generation is very much an open problem.

2) **Coordinating With Other Media:** When response generation is part of a larger interactive setting, including speech, graphics, and animation, as well as written language, a generator must coordinate its tasks with other components. For example, which information in the selected content should appear in language and which in graphics? If speech and animation are used, how are they to be coordinated temporally (e.g., how much can be said during a given scene)? What parameters used during response generation tasks should be made available to a speech component? These are issues that have only recently surfaced in the research community.

3) **Interaction Between Interpretation and Generation:** Many generation tasks use information sources that are also used for interpretation. How can sources be shared? For example, in order to provide responses that are sensitive to the user and to previous discourse, language generation needs access to a discourse history and a user model. While a history helps a response generator in determining what information can be left out and what terms to use, it helps an interpreter in resolving certain linguistic phenomena such as anaphoric reference. Both generation and interpretation need a lexicon and a grammar. While each has different needs, there is also overlap and duplication that can be avoided. In any of these tasks, there is a fine line between which uses of these knowledge sources fall into interpretation and which are part of generation. In the ideal case, interpretation and generation blend and certain components are used in both directions.

4) **Evaluating Generation Systems:** There has been very little work on how to measure success for a generation system. Possibilities include evaluating how well a user can complete a task which requires interaction with a system that generates responses, asking users to indicate satisfaction with system responses, performing a preference analysis between different types of text, degrading a response generation system and testing user satisfaction, and evaluating system generation against a target case. Each one of these has potential problems.

For example, task completion measures interact with the front end interface: that is, how easy it is for a user to request the information needed. Thus, it would be helpful to have interaction between computer scientists who build the systems and psychologists, who are better trained in creating valid evaluation techniques, to produce better ways for understanding how well a generation system works.

5) **Sources of Variability in Language Generation:** Natural languages allow for a wide range of variability in expressing information. While research in interpretation has often involved reducing different expressions to the same canonical form (e.g., active and passive forms are usually both converted to the same semantic representation ultimately), research in generation has often focused on identifying and representing constraints on language usage. If we can understand why different seemingly synonymous words are used in different situations, for example, we can understand when a generation system should select one word over another. Without such research, generation systems are forced to use random choice. Systems that rely too much on random choice often produce awkward and inappropriate language. This research has potential benefits for interpretation as well. Information about constraints on choice can provide information about the intent of the speaker when producing the utterance.

F. Speech Synthesis and Speech Generation

In human-computer interaction, the form of a computer response is as important as the content, and many applications require or are significantly enhanced by speech synthesis. The benefits are perhaps most clear in applications involving information access via telephone, computer training, and aids for the handicapped. In computer training, for example, research has shown that interactions via spoken responses resulted in better learning performance than visual presentation alone in a computerized course for teaching algebra [124], [125].

For remote access to computers via telephone, or for telephone information services, spoken responses are currently the only means of communication. Even for users interacting with computers locally, voice responses can reduce cognitive load in a multimedia environment or simplify an application with many response windows by providing a nonvisual information channel that can provide context for the visual information and help focus the user's attention.

Text-to-speech synthesis has applications for a broad array of problems, but is limited by the quality of current systems. In addition, advances in natural language generation open a new area of research, namely speech generation. Just as speech understanding involves more than simply sequencing speech recognition and natural language processing, so speech generation should involve more than simply connecting a response generation system to a text-to-speech synthesizer. Speech generation offers the potential for more natural speech synthesis, because the language generation process provides detailed semantic, syntactic, and dialogue information that can only be hypothesized in text-to-speech applications. In the context of speech generation, much work relating to focus and phrasing can be envisaged that was not previously possible

when text was the only input to synthesis. An additional new challenge is the coordination of understanding and response generation components in a spoken language system, particularly when the system assumes an active role in the dialogue.

Funding of speech synthesis research in the United States has lagged far behind funding of research in speech recognition and understanding. The reasons for this seem to be that 1) synthesis is thought to be a solved problem, or that 2) industry will fund the work. Speech synthesis is not a solved problem. Synthetic speech is not as intelligible or "acceptable" as natural speech, particularly for cases where language redundancy plays less of a role (e.g., in difficult material or unfamiliar names) or in lower quality audio environments [74]. The quality of current text-to-speech systems is a limiting factor in many applications, especially those where extensive output is required. As for industry funding, the results are not generally in the public domain, and consequently speech research has suffered. In contrast to a decade ago, it is difficult to gain access to a state-of-the-art synthesis system that will allow full control of the parameters necessary for conducting speech research.

Communication via spoken language involves two participants, the speaker and the hearer. If one participant's output capability is neglected, this leads to compromised and frustrating communication. Successful spoken language systems will only be possible if the system can both understand and speak intelligibly; speech synthesis and speech generation are technologies critical to this effort.

Key Research Challenges and Fundamental Scientific Issues: Many components involved in speech synthesis are common to both the text-to-speech and speech generation problems, and advances in basic speech synthesis algorithms will also advance speech generation. In fact, advances in basic speech synthesis technology may be critical to its effective use in human-computer interaction. Important research problems that should be addressed include:

1) Improvement in Basic Synthesis Technology: Of the many different components in a speech synthesis system which could be improved, a few particularly important research areas are: models of the physics of sound generation in the human vocal apparatus; models of articulation for synthesizing phonetic segments; theories of the relationship between prosody and syntax/semantics for predicting abstract prosodic patterns; and models of intonation and duration for interpreting those prosodic patterns acoustically [66], [27].

2) Computational Models of Variability: Explicit models of variability are needed in synthesis to avoid monotony, an issue both for synthesis of long monologues and long human-computer interactive sessions. In addition, models that can account for variability are more likely to also be useful in speech understanding applications, as demonstrated in [134] and [137].

3) Integration of Synthesis and Language Generation: Little work has been done on this problem, and there are many opportunities for exploiting the linguistic information that is a by-product of language generation. Possibilities range from simply increasing the quality to modeling discourse structure to active dialogue control.

4) Adaptation: Adaptation is an issue which is only recently being addressed [9]. Adaptation technology and, more generally, models that can be trained automatically, are important for adjusting a synthesis system to different situational demands, different speaker characteristics and style, and different languages, all of which will be important for more general applicability of speech synthesis. In particular, these methods are needed to handle systems that are very domain dependent or applications where there may be several modes of human-computer interaction.

5) Evaluation Metrics: As in other areas of speech and language research, the question of evaluation metrics needs to be addressed for speech synthesis. Current evaluation techniques address only segmental intelligibility, which is no longer the limiting factor in synthesis systems. Methods are now needed for evaluating systems at a higher level, e.g., in terms of cognitive load, naturalness and effectiveness in human-computer communication.

Speech synthesis provides an excellent domain in which to evaluate theories of speech communication, since the costs of speech synthesis experiments and system building are much lower than those for spoken language understanding. The same issues that appear to be missing in speech recognition are those that are missing in synthesis: accounting for variability in style, in dialect, in rate, determining the right units, combining them in a meaningful way, and so on. Putting effort into synthesis will pay off in terms of better quality synthesis as well as in better understanding of spoken language, which will in turn lead to improved models for recognition and understanding.

G. Multilingual Systems

Until recently, research in the United States in speech and natural language was almost exclusively aimed at monolingual communication in American English. However, cataclysmic shifts in geopolitics suggest that a reassessment of this approach is appropriate. The economy is increasingly global, from both the corporate and national perspectives. Military and diplomatic interests are creating increasing volumes of communication, and international telephone traffic is growing. The scientific community, though always somewhat international, is ever more so; as a result, data and published literature are larger and more multilingual. Moreover, advances in speech processing technology and the microelectronic technologies that support speech research have made a foray into multilingual systems feasible. This section outlines some of the issues in multilingual speech and language processing systems.

Key Research Challenges and Directions: A number of fundamental scientific issues must be addressed to arrive at a range of applications in multilingual speech and language processing. However, these scientific studies should be selected and guided by their relevance to a set of key research challenges in the field. These challenges include:

1) Multilingual Spoken Language Interfaces: Systems and techniques are needed which will allow users to speak to the systems in a variety of languages, and which will understand the speech well enough to efficiently carry out tasks such as interactive database retrieval [36] or command and control of complex systems.

2) Language Identification: As an independent capability or as a part of a multilingual spoken language system, techniques are needed to identify language and/or dialect in order to route the user to the appropriate human (e.g., human telecommunications operator) or automatic system (e.g., spoken language data retrieval system). A language identification system might utilize speech recognition techniques such as key word spotting, or language and speech recognition might operate jointly to both identify the language and recognize the spoken words.

3) Multilingual Text and Speech Generation from Multimodal Databases: Complementary to the multilingual input, techniques are needed to respond to the user in multiple languages, and to generate multiple forms of output (speech, text, video, graphics) in the language of the user.

4) Spoken Language Translation: This is the grandest of the challenges, encompassing all the above challenges plus a machine translation capability. Initial advances in this direction are indeed in progress [135], [106], [25], but considerable additional research will be necessary to achieve complete widely usable and robust speech translation systems. Short of completely fully automated translation, techniques are also needed to help human translators, by providing tools such as on-line dictionaries and grammars, and a mechanism for producing semi-automatic translation with interactive human review.

Fundamental Scientific Issues: A number of fundamental issues must be addressed to meet these challenges. Examples of such issues include:

- The general question of what are the fundamental acoustic, perceptual, and linguistic differences among languages should be investigated, with a view toward accommodating these differences in multilingual systems.
- An investigation should be undertaken of language-specific versus language-independent properties across languages. For example, is it possible to define language-independent acoustic/phonetic models, perhaps in terms of an interlingual acoustic/phonetic feature set?
- The innovation and evaluation of language-independent representations of meaning should be pursued, with a view toward the application of such representations in spoken language interfaces and/or spoken language translation systems.
- For spoken language translation, the fundamental issue of the granularity of translation should be addressed. What units (phrases, sentences, concepts) should be translated, and what is the effectiveness of literal translation versus paraphrasing. Some of these studies could be conducted using human translators executing a variety of controlled translation paradigms, including paradigms which accommodate the expected behavior of a speech understanding system feeding a speech synthesizer.
- Portability of spoken language system components needs to be studied. To what extent can system structures be language-independent, except for the use of language-specific training data and different vocabularies and grammars?
- In conjunction with the portability issue, formalisms and algorithms should be developed and studied for automatic learning and adaptation of spoken language representations at all linguistic levels (acoustic phonetics, prosody, syntax, semantics, pragmatics, and discourse), with the goal of facilitating multilingual applications.

H. Interactive Multimodal Systems

Multimodal systems could precipitate a major shift in the quality, utility, and accessibility of modern computing. They have the potential to support more flexible, easy to learn, and productive human-computer interactions. In addition, they are capable of producing more robust performance under adverse conditions, which in many cases will be required before spoken language technology can function adequately in realistic field environments. Multimodal systems also are expected to open up new and more challenging applications for computing, including interfaces for a new generation of portable computers. Since keyboards are incompatible with portability, interfaces to mobile computers necessarily must rely on input modalities like speech, handwriting, or direct manipulation, which are likely to be presented in multimodal combinations. We anticipate that multimodal systems, especially when situated on portable devices, will bring computing to a larger and more diverse user population than ever before.

Basic research is critically needed to guide the development of a new generation of multimodal systems. Advances in hardware speed and algorithms already are supporting the implementation of more transparent and natural communication modalities like spoken language, as well as the development of initial multimedia and multimodal systems. The aims of such systems include permitting people to speak and write in their own native language, to point or gesture while speaking, to view a synthesized human face with synchronized lip movements and emotional expressions while listening to speech, to participate in simulated virtual environments with accompanying speech, and to retrieve and manipulate information stored in rich multimedia formats (e.g., text, graphics, video, speech, hand drawn marks and writing, and so forth). However, the role that spoken language ultimately should play in future multimodal systems is not well understood [14]. In addition, since multimodal systems are relatively complex, the problem of how to design successful configurations is unlikely to be solved through a simple intuitive approach. Instead, determining optimal designs and appropriate applications for different types of multimodal systems will require interdisciplinary research, preferably based on advance simulations [94], [95].

There are many potential advantages of well-designed multimodal systems. One is the support of robust system performance under adverse conditions. For example, adequate recognition of spoken language could be maintained in a noisy environment with supplementary visual information about corresponding lip movements. The integration of visual and auditory information occurs naturally in face-to-face communication, with visual information gathered from the speaker's facial movements becoming relatively more salient in a noisy environment. Although contrasts like [b]/[d] and [m]/[n] are acoustically similar, and our ability to distinguish them is

degraded in a noisy environment, these contrasts nonetheless are easily distinguished when we observe a speaker's moving lips. In other cases, adequate recognition of spoken language could be supported with handwriting, graphics, or contextual information in virtual environments.

One clear experimental demonstration of how visual cues are integrated with auditory ones during speech recognition is provided by the "McGurk effect" [82]. During this effect, a person observes a videotaped face saying "ga" while listening to "ba" on a soundtrack. The perceptual result is that the auditory and visual information merges, such that the person reports hearing "da." Furthermore, the sound reported can be manipulated by having the person make judgments with eyes shut and open.

Inspired by these empirical results, computationalists have begun attempting to integrate auditory and visual information to improve the accuracy and robustness of speech recognition, with encouraging results [99], [100], [98], [145], [144]. For example, neural networks trained with combined visual and acoustic features have been shown to perform more accurately and degrade more gracefully as ambient noise levels are increased, compared to networks trained with acoustic features only [121]. Goldschen has obtained 25% recognition in continuous speech using optical recognition exclusively, without the use of acoustic data or syntax [38]. Such results support the belief that multimodal systems may display more desirable properties, especially under realistic field conditions, than stand-alone spoken language systems.

Apart from the issue of robustness, multimodal systems also offer the potential for broader utility, including the support of more challenging applications than those undertaken to date. For example, multimodal pen/voice systems aimed at the emerging mobile computing market could support a variety of new functions involving both computation and telecommunications, extending computational power to travelers, business and service people, students, and others working in field settings. Multimodal systems also could bring computing to a substantially larger and more diverse group of users than in the past. Examples include aged, disabled, and special populations whose specific sensory or intellectual limitations may be overcome by providing a choice among information channels, or merging sources of information from more than one channel.

Key Research Challenges and Fundamental Scientific Issues: In order to work toward the development of more facile and productive multimodal systems, many key research challenges and fundamental scientific issues will need to be addressed. Among these challenges are the following:

1) Performance Characteristics of Multimodal Systems: Interdisciplinary research will be needed to design multimodal systems with performance characteristics superior to those of simpler unimodal alternatives. This will require empirical work with human subjects, the construction of new prototype systems, and the development of appropriate metrics for evaluating the accuracy, efficiency, learnability, and expressive power of different multimodal systems.

2) Coordination Among Modalities: Strategies will be needed for coordinating input and output modalities, and for resolving integration and synchronization issues among the modalities functioning during input and output. For example, the ability to use information from one input modality to disambiguate simultaneous input from another will be required.

3) Component Technologies: More research will be needed to develop newly emerging component technologies that are required to build multimodal systems, such as spoken language recognition, handwriting recognition and integrated pen systems, natural language processing, gesture recognition, 3-D virtual reality and its various sensory components, technology for assessing human gaze patterns, technology for simulating lip movements and expressions on the human face, and so forth. Priority should be given to supporting the more promising but underdeveloped component technologies, in the light of successful developments in multimodal systems.

4) Theory of Communication Modalities: In order to build principled multimodal systems, a better understanding will be required of the unique structural, linguistic, and performance characteristics of individual communication modalities, as well as properties associated with interactions among modalities. From this foundation of information, comprehensive theoretical models need to be constructed from which predictions can be made about the strengths, weaknesses, and overall performance of different types of unimodal and multimodal systems.

5) General Treatment of Multimodal Dialogue: A general theory of communicative interaction will be needed to provide a foundation for handling interactive dialogue in a manner that is independent of the specific input and output modalities used in any given multimodal system. Such a theoretical approach would provide the basis for implementing a successful coordination among the different modalities in the multimodal system.

6) Research Methodology and Evaluation: Since multimodal systems represent hybrid communication forms, often without natural analogues, there is a special need for better simulation tools to collect advance data on people's language and performance in different simulated multimodal arrangements, so that systems can be designed accordingly. New simulation methods will have to be devised to accommodate the different component technologies represented in planned multimodal systems. In addition, appropriate methods are needed for scientifically evaluating the performance of multimodal systems.

III. INFRASTRUCTURE

Spoken language processing is a field where it is particularly important to support the scientific infrastructure. The problems are inherently multidisciplinary, and infrastructure supporting communication between researchers is invaluable. It is imperative that investigators working at different sites be able to cooperate and exchange data and software across sites. The field also requires an infrastructure for training researchers with the necessary skills, and for providing the computer resources, algorithms, data, and tools needed to optimize productivity.

A. Multidisciplinary Research and Training

Research in spoken language understanding often requires expertise in diverse areas such as speech and hearing science, linguistics, psychology, signal processing, statistics, pattern recognition, and computer science. The multidisciplinary nature of the research makes it unlikely that a single research group can conduct meaningful research across the entire spectrum. As a result, we must encourage collaborative research.

The multidisciplinary nature of spoken language research also means that it does not fit well into the department structure of a university. There are relatively few universities that train researchers in computational linguistics or speech recognition, and fewer still that train researchers in the area of spoken language understanding.

To fill this gap, it is necessary to create multidisciplinary academic programs in spoken language systems at our educational institutions, as well as a summer Spoken Language Institute. There is clear need for such a program: Victor Zue and the MIT Spoken Language Group have offered a week-long spectrogram reading course every other summer for the past six years, with a full enrollment each time. For a Spoken Language Institute, the program might consist of several such week-long intensive courses in core areas such as spectrogram reading, speech recognition by human and machine, dialogue modeling and understanding of spontaneous speech. This might also be the appropriate setting for workshops (e.g., on prosody and prosodic annotation) and mini-courses on other topics such as speech synthesis, language generation, or a course in the use of basic speech tools. Such a summer program would fill a significant gap in training young researchers, as well as supporting cross-training of established researchers and bringing together researchers from many disciplines to facilitate collaboration.

The need for multidisciplinary collaboration is particularly clear in the area of multimodal systems. Cross-training of researchers is clearly critical in this area, but not sufficient. Research progress on multimodal systems is most likely to be accomplished through a combination of innovative empirical and computational work, ideally conducted by well-coordinated interdisciplinary teams. In addition, such teams either must include or have close access to expertise representing the technologies incorporated in the multimodal system under study or development. In practice, this often may require close working relations between basic researchers in academics or research institutes and engineers in industrial settings who are developing core technologies and applied systems. Such considerations also apply to the development of multilingual systems, where researchers fluent in the appropriate languages are required for system development. Since such teams must represent a span of disciplines, technologies, research sites, and even countries, they may frequently require relatively large working groups.

B. Corpus Development, Evaluation, and Resource Sharing

The availability of common corpora of speech and text is a critical resource that has been partly responsible for the significant gains made in speech and language processing in recent years. These corpora have also been associated with standardized evaluations of the component technologies [96], [24]. Although the use of regular "common evaluations" originated in the ARPA community, it has spread to the broader community and provides periodic measurements of progress of the field over time, as well as making it possible for individual systems to evaluate their internal progress. The corpora and evaluation sites require mechanisms for distribution, so that sites that wish to obtain data or to participate in evaluations can obtain the necessary information and software. The groundwork for these three ingredients—corpus development, evaluation metrics, and resource sharing mechanisms—has been put into place for the spoken language community, but we must continue to support this infrastructure and extend it to meet new research objectives.

Corpus Development: Large amounts of speech data spanning the range from highly focused tasks to unconstrained conversations are needed to model the many sources of variability described in this paper. These data are necessary both to develop the statistical models and theoretical foundations for language representations. The scientific community requires: a) timely access to these data; b) sufficient computer resources to store and process the data; and c) speech research tools to display and process the data.

There are a number of important unresolved issues in corpus development, including transcription conventions, levels of description (sounds, words), and a reliable system for transcribing the prosodic structure of speech. Continuation of these efforts requires support for workshops to define the research community's need, continuing support for data collection, and support to provide the necessary computer resources and tools.

Data collection and related infrastructure support must be extended to research in speech synthesis. There is an important need for data collection efforts in support of speech synthesis research, for the development of evaluation criteria for speech synthesis, and for the development of procedures for communicating about and sharing the data.

Evaluation Methodology: Evaluation has played a central role in the ARPA Spoken Language program. Evaluation methods are now in place to evaluate speech recognition (in terms of word accuracy), language understanding (in terms of retrieval of the correct database answer, given a transcription of the spoken input), and spoken language (also in terms of the correct database answer, given the speech). This methodology provides an automated evaluation for prerecorded (speech or transcribed) data, allowing sites to iteratively train and evaluate their systems (in the limited domain of air travel planning). The availability of a significant corpus of transcribed and annotated training data (14 000 utterances of speech data, with 7500 utterances annotated with their correct answer) has provided an infrastructure leading to very rapid progress in spoken language understanding.

As spoken language systems become more sophisticated (and more interactive), the research community will need new ways of evaluating its systems. Evaluation methodologies, in turn, will drive new data collection efforts and create new software resources to be shared. Research into appropriate

evaluation methods is of critical importance, to ensure that systems are optimized according to relevant criteria.

Resource Sharing: The collection, transcription, and distribution of speech data for spoken language understanding is a massive task. In the U.S., ARPA has supported speech data collection and distribution of these data to the research community through NIST. Recently, as a precompetitive initiative, the U.S. Congress has established the Linguistic Data Consortium, administered through ARPA, to expand this activity. At the Oregon Graduate Institute, the Center for Spoken Language Understanding collects and distributes multilanguage telephone speech corpora to universities free of charge.

At the international level, the Cocosda committee meets at international meetings to promote communication and awareness of speech corpora, as well as standardization and sharing. This infrastructure is critical to the continued development of spoken language systems. It is facilitated by commonality of environment (UNIX) and language (e.g., C), but also requires increased investment at individual sites, to provide the necessary mass storage capacity and computing facilities to use the data and software.

C. Computational Resources

The majority of speech recognition research is most productively conducted using local high-speed workstations. With the rapid advances in computer technology, the price/performance ratio continues to improve by roughly a factor of two each year. Consequently, workstations should be replaced, on the average every two to three years—about the length of a typical research grant. Institutional support, particularly at universities, is often not sufficient to allow this level of replacement. More support for workstation replacement would be a cost-effective investment to improve research productivity. Some degree of hardware/software standardization would also be useful to promote better sharing of software and algorithms among various sites, including the capability to capture and playback speech. The following needs are apparent:

High Performance—Computation, Memory Bandwidth, and Storage: The tasks we are undertaking today are computationally very demanding. The lessons of the recent history of research in speech and natural language processing show that whatever computational power is available, it can always be used, and then some.

There are two reasons for our needs for high performance computing. First, there is a real-time constraint which limits the complexity of the spoken language processing for any given hardware capability. A second, more general point is that noninteractive computing still requires the analysis of huge corpora, for instance, for the training of a speaker-independent recognizer. If it takes too long to do the analysis, then the experiment will simply not be attempted. This is obviously an impediment to progress. Our choice of experimental algorithms is now restricted because of computational efficiency, so that the runs can be done in reasonable time. It would be preferable to broaden the research to include more complex approaches. Some emerging technologies, such

as connectionist networks and more detailed models of the physics of auditory and vocal tract mechanisms, require much more computation than current mainstream techniques—but it is just this kind of nontraditional approach that needs to be funded.

In addition to fast arithmetic (for experimental algorithms) and massive storage (for large corpora), computational systems for this purpose require significant memory bandwidth. Without this feature, fast arithmetic units are starved for data. Unfortunately, standard caching schemes frequently are not sufficient for this purpose with speech problems. Therefore, more specialized architectures are frequently considered for speech processing purposes.

General Purpose versus Special Architectures: Both general purpose and special purpose machines are useful in speech research. The general purpose machines are usually easier to program than their special purpose counterparts. This saves costly human labor. Since researchers rarely use a program more than a few times before modification, ease of programming is paramount.

There are some cases, however, in which raw computing speed is essential. Recent advances have made it possible to do near-real time systems without special purpose hardware—for the "standard" approaches, such as hidden Markov models. However, if the experiment involves new, computationally intensive algorithms such as segment-based modeling or fluid-dynamic modeling for speech recognition, it is often necessary to use a special architecture designed expressly for the experimental purpose. Such systems are quite difficult to program but, once programmed, yield significantly higher execution speeds. As noted above, frequently the most important characteristic of the special purpose processor (from the standpoint of speech processing) is the facility for data movement to the arithmetic units.

Real Time I/O and Networking at Audio Data Rates: In many laboratories, computing power is inexpensively achieved by networking many workstations together. It is also the case that remote laboratories may wish to conduct a common experiment by linking their respective computers with a network. In either case, the network is an intrinsic bottleneck in attempting any real-time on-line experiments. While the data rates for speech (typically 8 to 32 KB/sec) are not prohibitive for Ethernet technology, these networks do not provide a real-time guarantee, so that interactive experiments may experience nondeterministic delays. For this reason it is important for speech researchers to keep abreast of newer network technologies.

Optimizing Compilers for Parallel Architectures: Compared with general purpose scalar machines, current parallel and application specific architectures typically require considerable programmer effort to produce efficient code. A researcher, who is typically not an expert programmer, requires the support of an optimizing compiler to help obtain maximum performance from a machine. Compilers for specific parallel machines are already quite competent at discovering and exploiting some forms of parallelism; for instance, vectorizing compilers can successfully execute most of the operations in many loop nests in vector mode. However, efficiently mapping an arbitrary

piece of code to an arbitrary parallel architecture is beyond the reach of current compilers.

More development is needed in this area, both in expanding the capabilities of compilers for conventional languages such as FORTRAN and C/C++, and in developing newer languages which are more suitable for exposing the parallelism in application codes. In the meanwhile, the use of specialized, hand-crafted libraries for common tasks can be used to give researchers access to the power of novel architectures.

D. Sharing of Speech Research Tools and Algorithms

The development of an integrated set of speech tools is an important priority. Speech research requires software to generate and display signal representations, to edit, label, and listen to speech, to train classifiers and build working systems. The development of speech research tools and useful algorithms (e.g., Viterbi search) is labor intensive and often redundant across laboratories. Commercially available speech tools are available, but they are expensive.

In the U.S., as part of its Software Capitalization Program, NSF has funded the development of a set of portable tools for speech recognition research for UNIX workstations running X windows. These tools [28] are available to interested researchers[3]. Similarly, ARPA has funded an effort to make standard speech recognition components available, through Entropic Research Laboratory. If these programs are successful, they should be expanded to provide software support for other areas of spoken language understanding.

E. Communication

In recent years, there have been changes in the dissemination of knowledge that have affected research in speech and natural language processing. With large-scale research efforts and ever faster change in knowledge, algorithms, and techniques, advances in speech and language research are increasingly communicated through direct contact, workshops, conferences, and electronic mail exchange among partners of major research projects rather than through major journals, books, and publications. This fact has the disadvantage that it encourages the formation of "inside player cliques" and makes it increasingly difficult for smaller laboratories or single researchers to participate.

More effective communication can be promoted by supporting infrastructure for easy access to and rapid exchange of information among interested researchers. This infrastructure can be accomplished in a number of ways:

1) A Publications Database Facility for Unreviewed Preprints, Publications Workshop Summaries and Protocols for General Review and Consumption: An example for such a facility already in existence is the Neuroprose archives maintained by Jordan Pollack at Ohio State. Individuals may deposit reviewed or unreviewed research reports and technical notes/reports to this database, allowing other researchers access to these reports via anonymous FTP over the network.

[3] Details regarding the OGI speech tools and how to get them are available via anonymous ftp from **speech.cse.ogi.edu**. Change directory to the /pub/tools sub-directory and get the file entitled **ANNOUNCE**.

Results reported here have of course not been reviewed and must be accepted with caution, but this allows for rapid dissemination of ideas and preliminary results much like conferences, workshops, and personal communication would. Such a facility is not only efficient, but saves the costs for shipping and handling as well (the recipient prints it on his/her own printer). Such a database is self-maintaining and requires minimal attention and resources.

2) Facility for Electronic Benchmarking and Other Algorithms: Such a facility would be a repository for common corpora, algorithms and tasks, for quick and easy access by researchers and for evaluation purposes. Performance results could be published there as well. Very similar to the above-mentioned depositories, this facility should require minimal maintenance, although it might require fast datalinks, when databases become excessively large.

3) Video-Conferencing, Multimedia Facilities: These facilities, perhaps integrated on turnkey workstations, could significantly improve the personal communication between researchers in distributed locations working on big projects jointly. In particular, face-to-face communication via multimedia windows, joint access to shared reference material, and joint manipulation and annotation of such material should be explored. Transmission of on-line sketches and drawings could be helpful as well. Such facilities might someday also enable distributed teams of researchers to work together on joint projects under joint funding.

IV. BENEFITS

Spoken language systems have the potential to make on-line information resources readily available to vast new classes of users—hands-busy and eyes-busy users, casual users, novices, handicapped users—by providing a convenient and natural modality (speech) to access and manipulate information. They provide the ability to support international cooperation, diplomacy, and commerce in the increasingly interconnected global economy via multilingual and speech-to-speech translation systems. In this section, we examine some of the benefits of spoken language systems as well as the benefits of supporting research and education in this area.

A. Societal Impact of Spoken Language Systems

The spin-offs of academic, governmental, and industrial research on the production, perception, recognition, and understanding of speech are many and far-reaching. The benefits are most likely to appear first in three general areas: hands-busy/eyes-busy applications; aids for the disabled; and increased access to on-line information resources for the general public.

Spoken language systems should increase productivity. The majority of the population is neither computer literate nor trained in typing. While continuous speech ranges from 150 to 250 words per minute, a trained typist averages only 60 words and most of us type much more slowly. So, with speech input, we would expect input productivity improvements ranging from at least 3 to 15 times, in addition to the ability of perform-

ing tasks which would not be possible at all if the operator's hands had to be devoted to the keyboard. This is particularly important in some critical military and civilian activities where the hands of a pilot, surgeon, or machinery operator are used in vital, time-critical tasks, while voice commands could be recognized and carried out. Spoken language technology will also enable a new set of hand-held, highly portable devices which cannot accommodate keyboards.

People who suffer from various disabilities sometimes find it difficult to control their environments. Such individuals can be helped by speech technology to a greater extent than others because they are often deprived completely of some needed capability; thus, any help offered by technology greatly enhances their quality of life. Moreover, such individuals will be highly motivated to use the technology, since it affords them abilities they would otherwise be without. Such motivation can often overcome shortcomings of existing speech processing devices. The main applications of speech recognition in this connection are those of controlling machines by voice. This includes not only everyday appliances such as televisions and room lights, but also personal devices such as wheelchairs and special beds. Similarly, speech synthesis affords the capability to communicate by voice to those who have lost their natural ability to do so. For such individuals, speech synthesizers can be operated by any number of means from typing to pointing to icons with a mouth-held stick or even by eye-tracking. Some primitive devices of this type are already commercially available. For the visually impaired, recognition of printed characters, combined with speech synthesis, provides a natural way to present ordinary text.

Spoken language systems will increase the public's access to information, from travel schedules to emergency medical information. Even people who have access to computers via PCs and modems often have no idea how to retrieve pertinent information. Thus many important sources of information are still difficult for people to access because of unavailability of terminals and because of the need to use arcane programming languages to access that information. There is, however, a universally available terminal which everyone knows how to operate the telephone. Speech recognition and synthesis can provide universal, simple access to machine-readable databases via commands spoken over the telephone. Again, rudimentary devices providing such a capability are beginning to appear commercially. However, the real benefits will be realized when the transactions can be conducted via a natural, colloquial discourse.

Speech understanding and conversational systems are within the scope of our current research agenda. The benefits of such technology would improve the productivity of the technical and specialized users of computers, but equally important, it opens the possibility of using computers, telecommunication and transaction handling equipment, messaging, and mechanical actuators in general, to the wider population at large who may not be computer literate or scientifically minded. Enabling computers to recognize and understand speech would also be a major boom to the computer industry because of the enormous variety of new applications in which these machines could be useful.

B. Commerce

Spoken language systems are among the many technologies that will make a significant contribution to cost reduction and new revenue generation in industry today. The computer, financial, entertainment, and telecommunications industries have made it clear that this technology is important to their future competitive positions by forming major research and development organizations devoted to speech systems.

Easy information access and control are key to each of these industries. Speech understanding systems allow ordinary telephones to act as database access terminals. Multibillion-dollar customer service operations can be equipped with spoken language interfaces to reduce costs and offer new services at a competitive price. Office and factory automation will make companies run more efficiently and more cost effectively. In addition, the National Information Infrastructure is making electronic, network-based communications and commerce available to vast new market segments. The need for advanced interfaces, including spoken language interfaces, is already apparent, as people struggle to utilize this new information source effectively.

C. International Cooperation and Business

Multilingual speech and language processing could have a major impact on the economic future of our society. With increasing internationalization, it becomes exceedingly important for individuals to communicate and cooperate with colleagues, offices, laboratories, and customers in other countries. Bridging language barriers could open yet untapped markets and avenues for trade. Large corporations also stand to benefit: many have world-wide networks of laboratories and sales offices. Cross-language collaboration is becoming more and more commonplace. Overcoming communication barriers will decrease costs and improve productivity.

Beyond economic advantages in trade, sales, management, and engineering, it is also becoming increasingly important to stay abreast of scientific and economic developments that may not be readily available in English. Expanding the available body of knowledge by tapping foreign information rapidly could lead to strategic advantages in these areas. Similarly, efficient access to such information may also be of vital importance to our security needs and diplomatic efforts.

Understanding and overcoming language barriers may improve interaction among the many peoples within our society. It may also raise an awareness of other cultures and languages, and we would hope to see a new generation of multilingually aware and adept students emerge as a side effect. Finally, multilingual speech and language processing may aid in this educational process, providing easily accessible computer-aided language instruction.

D. Benefit to Scientific Community

Spoken language systems will increase the productivity of researchers who work with computers on a daily basis. As multilingual spoken language systems become available, the handicap of language differences between researchers will vanish and many more possibilities for international collaboration

will exist. Moreover, automatic recognition and translation systems may be well-suited to being tested first within a scientific community with a strong need for international communication.

The field of computer science itself can evolve quite markedly as human speech and language become an alternative to and, in some cases, replace current human-computer interfaces. For example, programming interfaces can, with appropriate design of speech input systems, be made less constrained than keyboard entry systems. Spoken language systems can do away with the need for a full keyboard, thus making palm-top computer technology more viable. In summary, the evolution and near-term performance improvements of speech and language systems will have a major impact within the scientific and engineering disciplines, as well as within the general population.

E. Student Education and Jobs

Research and educational activity in spoken language understanding is an essential part of the infrastructure for a healthy industry. The research provides the theoretical foundation and discoveries for new technology, it provides the future leaders of the industry, and it educates those who work in it.

The need for trained researchers in spoken language systems is already apparent. Well-trained researchers are rare and in great demand; for example, several start-up efforts are now waiting for qualified leaders. There are few multidisciplinary centers of excellence in spoken language understanding worldwide with multidisciplinary curricula, and competition for the few graduating students is intense. By increasing support for graduate students, by supporting cross training of researchers in related disciplines (e.g., signal processing, linguistics), and by providing support for programs such as a Summer Spoken Language Institute, the pool of researchers can be enlarged to meet the increasing demand. Students trained in the field of spoken language processing will then have the valuable experience of working in a multidisciplinary field, where they will develop the technical and the collaboration skills needed to solve complex problems wherever they end up working: academia, industry, or government.

V. Conclusion

Spoken language systems have reached a critical point: Research has made breakthroughs in the last decade and the technology is now poised on the threshold of usability. With a steady increase in computing power available, the growing reliance on rapid electronic communication, and the maturing of spoken language technologies, we can expect that we will see functional, commercially viable spoken language interfaces emerge before the year 2000. We are already seeing the beginnings of this process: small vocabulary recognition for automated telephone call handling, routine packaging of (low-end) recognizers and synthesizers with workstations, and phone-based speaker identification applications.

In the near future, we can expect to see more sophisticated applications, such as limited menu and forms-based phone information access (automated banking, request for timetables).

These will lay the groundwork for more sophisticated systems involving continuous speech recognition and sophisticated language processing technology beyond just simple phrase recognition.

The 1990s will be a critical decade for spoken language research. Recent technological advances have resulted in market forces that create a dramatic need for spoken language interfaces. The cellular telecommunications market is the fastest growing market in history, with sales of approximately 20 000 cellular phones per day. These portable devices will serve as gateways to vast networks providing information and services. Many of today's personal and business transactions will be conducted over these networks through human-computer dialogues—providing we meet the research challenges described in this article.

To bring the technology to fruition, we will have to make a major investment—spoken language technology is inherently a cross-disciplinary technology which does not fit neatly into an academic department framework. In order to continue making progress in research and to supply the rapidly increasing demand from industry, we need to encourage researchers to reach across the boundaries of narrowly defined fields, and we need to support students, making infrastructure, equipment and educational resources available to them.

References

[1] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. IEEE Int. Conf. ICASSP*, 1990, pp. 849–852.
[2] J. Allen *et al.*, "The discourse system project," Comput. Sci. Dept., Univ. of Rochester, Tech. Rep. 317, Nov. 1989.
[3] J. F. Allen and C. R. Perrault, "Analyzing intention in utterances," *Artif. Intell.*, vol. 3, no. 15, pp. 143–178, 1980.
[4] D. Appelt and E. Jackson, "SRI International February 1992 ATIS benchmark test results," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.
[5] M. Bates *et al.*, "The BBN/HARC spoken language understanding system," in *Proc. 1993 IEEE ICASSP*.
[6] M. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction (SRC)," *IEEE Trans. Signal Processing*, vol. 39, no. 11, pp. 2377–2386, 1991.
[7] G. Brown, K. Currie, and J. Kenworthy, *Questions of Intonation*. Croom Helm, 1980.
[8] B. Butterworth, "Hesitation and semantic planning in speech," *J. of Psycholinguistic Res.*, vol. 4, pp. 75–87, 1975.
[9] R. Carlson and B. Granstrom, "Speech synthesis development and phonetic research—A personal introduction," *J. Phonetics*, vol. 19, pp. 3–8, 1991.
[10] F. Chen and M. Withgott, "The use of emphasis to automatically summarize a spoken discourse," in *Proc. 1992 ICASSP*, pp. I.229–232, Mar. 1992.
[11] J. Cheshire, *English Around the World: Sociolinguistic Perspectives*. Cambridge, UK: Cambridge Univ. Press, 1991.
[12] L. A. Chistovich and V. V. Lublinskaya, "The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli," *Hearing Res.*, vol. 1, pp. 185–195, 1979.
[13] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2623–2629, 1989.

[14] P. R. Cohen and S. L. Oviatt, *The Role of Voice in Human-Machine Communication*. Washington, DC: National Academy Press, 1994, Chap. 1.

[15] D. V. Compernolle, *et al.*, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communicat.*, vol. 9, no. 5/6, p. 433–442, Dec. 1990.

[16] N. Daly and V. Zue, "Statistical and linguistic analyses of F0 in read and spontaneous speech," in *Proc. Int. Conf. Spoken Language Processing*, Oct. 1992, pp. I:763–766.

[17] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds," in *Proc. 1994 IEEE ICASSP* (Adelaide, Australia), Apr. 1994, pp. I:45–48.

[18] J. DiPaolo and A. Faber, "Phonation differences and the phonetic context of the tense-lax contrast in Utah English," *Lang. Variation and Change*, vol. 2, pp. 155–204, 1991.

[19] J. Edwards, M. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *J. Acoust. Soc. Amer.*, vol. 89, pp. 369–382, 1991.

[20] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, no. 6, pp. 1303–1316, June 1992.

[21] A. Erell and M. Weintraub, "Recognition of noisy speech: Using minimum-mean log-spectral distance estimation," in *DARPA Workshop Spoken Lang. Syst.*, June 1990, pp. 341–345.

[22] L. D. Erman *et al.*, "The Hearsay II speech understanding system," *ACM Computing Surveys*, pp. 213–253, 1980.

[23] L. Hirschman *et al.*, "Multi-site data collection for a spoken language corpus," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.

[24] L. Hirschman *et al.*, "Multi-site data collection and evaluation in spoken language understanding," in *Proc. Human Lang. Technol. Workshop* (Princeton, NJ), Mar. 1993.

[25] M. Rayner *et al.*, "Spoken language translation with mid-90's technology: A case study," in *Proc. 3rd European Conf. Speech Commun., Technol.* (Berlin), Sept. 1993.

[26] W. Ward *et al.*, "Speech recognition in open tasks," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.

[27] G. Fant, "What can basic research contribute to speech synthesis?" *J. Phonetics*, vol. 19, pp. 75–90, 1991.

[28] M. Fanty, J. Pochmara, and R. A. Cole, "An interactive environment for speech recognition research," in *Proc. Int. Conf. Spoken Language Processing* (Banff, Alb., Canada), Oct. 12–16, 1992.

[29] T. W. Finin, "GUMS: A general user modelling shell," in *User Models in Dialogue Systems*, A. Kobsa and W. Wahlster, Eds. New York: Springer-Verlag, 1989, pp. 411–430.

[30] J. L. Flanagan, "Use of acoustic filtering to control the beamwidth of steered microphone arrays," *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 423–428, Aug. 1985.

[31] J. L. Flanagan *et al.*, "Autodirective microphone systems," *Acoustica*, vol. 73, pp. 58–71, Feb. 1991.

[32] J. E. Flege, "Laryngeal timing and phonation onset in utterance-initial English stops," *J. Phonetics*, vol. 10, pp. 177–192, 1982.

[33] T. Gay, "Mechanisms in the control of speech rate," *Phonetica*, vol. 38, pp. 148–158, 1981.

[34] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment," *J. Phonetics*, vol. 16, no. 1, pp. 109–124, 1988.

[35] H. Gish, Y. L. Chow, and J. R. Rohlicek, "Probabilistic vector mapping of noisy speech parameters for HMM word spotting," in *Proc. Int. 1990 IEEE ICASSP*, pp. 117–120.

[36] J. Glass *et al.*, "A bilingual voyager system," in *Workshop Human Lang. Technol.*, Mar. 1993.

[37] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. 1992 IEEE ICASSP*, pp. 517–520.

[38] A. J. Goldschen, "Continuous automatic speech recognition by lipreading," Ph.D. thesis, George Washington Univ., Sept. 1993.

[39] B. J. Grosz, "Discourse analysis," in *Understanding Spoken Language*, D. E. Walker, Ed. New York: North Holland/Elsevier, 1978, pp. 235–268.

[40] B. J. Grosz and C. L. Sidner, "Attentions, intentions, and the structure of discourse," *Computat. Linguistics*, vol. 3, no. 12, pp. 175–204, 1986.

[41] G. Guy, "Variation in the group and in the individual: The case of final stop deletion," in *Locating Language in Time and Space*, W. Labov, Ed. New York: Academic, 1980.

[42] M. A. K. Halliday, *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. Baltimore: University Park Press, 1978.

[43] S. Hamlet, "Handedness and articulatory asymmetries in /s/ and /l/," *J. Phonetics*, vol. 15, pp. 191–195, 1987.

[44] J. Hampshire and A. Waibel, "The meta-pi network: Connectionist rapid adaptation for high-performance multi-speaker phoneme recognition," in *IEEE ICASSP*, Apr. 1990.

[45] B.A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35. no. 7, pp. 968–973, 1987.

[46] H. Hermansky, "Perceptual linear predictive PLP analysis for speech," *J. Amer. Acoust. Soc.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[47] H. Hermansky *et al.*, "Compensation for the effects of the communication channel in auditory-like analysis of speech," in *Proc. 2nd European Conf. Speech, Commun., Technol.* (Genova, Italy), Sept. 1991, pp. 1367–1370.

[48] _____, "RASTA-PLP speech analysis technique," in *Proc. 1992 IEEE ICASSP*, Mar. 1992, pp. I:121–124.

[49] H. Hermansky and D. J. Broad, "The effective second formant f2' and the vocal tract front-cavity," in *Proc. 1989 IEEE ICASSP*, pp. 480–483.

[50] H. Hermansky, N. Morgan, and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on rasta processing," in *Proc. 1993 IEEE ICASSP,*, pp. II:83–86.

[51] D. Hindle, "The social and situational conditioning of phonetic variation," Ph.D. thesis, Univ. of Pennsylvania, 1980.

[52] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. 2nd European Conf. Speech Commun., Technol.* (Genova, Italy), Sept. 1991, pp. 413–416.

[53] J. Hirschberg, "Towards a redefinition of yes/no question," in *22nd Annual Meeting Assoc. for Computat. Linguistics* (Stanford University, Stanford, CA), 1983.

[54] J. Hirschberg and B. Grosz, "Intonational features of local and global discourse," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.

[55] _____, "Intonational features of local and global discourse structure," in *Proc. 5th DARPA Workshop Speech, Natural Lang.*, Feb. 1992.

[56] J. Hirschberg and G. Ward, "The influence of pitch range, duration, amplitude, and spectral features on the interpretation of rise-fall-rise intonation contour in English," *J. Phonetics*, vol. 20, pp. 241–251, 1992.

[57] L. Hirschman and C. Pao, "The cost of errors in a spoken language system," in *Proc. Eurospeech '93* (Berlin), Sept. 1993, vol. 2, pp. 1419–1422.

[58] H. W. Hon and K. F. Lee, "Vocabulary learning and environment normalization in vocabulary-independent speech recognition," in *Proc. 1992 IEEE ICASSP*, Mar. 1992, pp. I-485–488.

[59] A. Hughes and P. Trudgill, *English Accents and Dialects: An Introduction to Social and Regional Varieties of British English*. London: Arnold, 1979.

[60] K. I. Iso, "Speech recognition using dynamical model of speech production," in *Proc. 1993 IEEE ICASSP* (Minneapolis, MN), Apr. 1993, pp. II:283–286.

[61] K. Johnson, "The role of perceived speaker identity in F0 normalization of vowels," *J. Acoust. Soc. Amer.*, vol. 88, pp. 642–654, 1990.

[62] B. H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," in *Proc. 1987 IEEE ICASSP*, pp. 2368–2371.

[63] S. J. Kaplan, "Cooperative responses from a portable natural language query system," *Artif. Intell.*, vol. 19, no. 2, 1982.

[64] I. Karlsson, "Female voices in speech synthesis," *J. Phonetics*, vol. 19, pp. 111–120, 1991.

[65] D. Klatt and L. Klatt, "Analysis, synthesis and perception of voice quality variations among male and female talkers," *J. Acoust. Soc. Amer.*, vol. 87, pp. 820–857, 1990.

[66] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, 1987.

[67] A. Kobsa and E. W. Wahlster, *User Models in Dialogue Systems*. New York: Springer-Verlag, 1989.

[68] R. Kuhn and R. DeMori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 6, pp. 570–583, June 1990.

[69] W. Labov, *Language in the Inner City; Studies in the Black English Vernacular*. Philadelphia: Univ. of Pennsylvania Press, 1972.

[70] _____, "Sources of inherent variation in speech," in *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Erlbaum Associates, 1986, pp. 402–423.

[71] W. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, p. 41–104, 1983.

[72] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[73] C. Linde and J. Goguen, "Structure of planning discourse," *J. Social Biol. Structure*, vol. 1, pp. 219–251, 1978.

[74] J. S. Logan, B. G. Greene, and D. B. Pisoni, "Segmental intelligibility of synthetic speech produced by ten text-to-speech systems," *J. Acoust. Soc. Amer.*, vol. 86, pp. 566–581, 1986.

[75] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 7, pp. 1119–1134, 1988.

[76] D. D. MacDonald, "Natural language generation as a computational problem: An introduction," in *Computational Models of Discourse*, M. Brady and R. C. Berwick, Eds. Cambridge, MA: M.I.T. Press, 1983.

[77] A. Malhotra, "Design criteria for a knowledge-based English language system for management: An experimental analysis," Mass. Inst. Technol., Tech. Rep. MAC TR-146, 1975.

[78] W. C. Mann and J. A. Moore, "Computer generation of multiparagraph English text," *Amer. J. Computational Linguistics*, vol. 1, no. 7, 1981.

[79] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," in *Proc. 1988 IEEE ICASSP*, pp. 36–39.

[80] K. F. McCoy, "The ROMPER system: Responding to object-related misconceptions using perspective," in *24th Annual Meeting Assoc. Computational Linguistics* (New York, NY), June 1986.

[81] M. McCutcheon, A. Hasegawa, and S. Fletcher, "Effects of palatal morphology on [s, z] articulation," *J. Acoust. Soc. Amer.*, vol. 67, pp. S94, 1980.

[82] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[83] K. R. McKeown, *Text Generation*. Cambridge, UK: Cambridge Univ. Press, 1985.

[84] K. R. McKeown, *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge, UK: Cambridge University Press, 1985.

[85] K. R. McKeown and W. R. Swartout, "Language generation and explanation," in *Annual Review of Computer Science*, J. F. Traub *et al.*, Ed. Palo Alto, CA: Annual Reviews Inc., 1987.

[86] G. A. Miller *et al.*, "A semantic concordance," in *Proc. ARPA Human Lang. Technol. Workshop* (Princeton, NJ), Mar. 1993.

[87] J. M. Mullenix, D. B. Pisoni, and C. S. Martin, "Some effects of talker variability on spoken word recognition," *J. Acoust. Soc. Amer.*, vol. 85, pp. 365–378, 1989.

[88] A. Nadas, D. Nahamoo, and M. Picheny, "Adaptive labeling: Normalization of speech by adaptive transformations based on vector quantization," in *Proc. 1988 IEEE ICASSP*, pp. 521–524.

[89] _____, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[90] T. M. Neary, "Static, dynamic and relational properties in vowel perception," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2088–2113, 1989.

[91] F. Nolan, *The Phonetic Basis of Speaker Recognition*. Cambridge, UK: Cambridge Univ. Press, 1983.

[92] F. Nolan and P. E. Kerswill, "The description of connected speech processes," in *Studies in the Pronunciation of English: A Commemorative Volume in Honour of A. C. Gibson*, S. Ramsaran, Ed. London: Routledge, 1990.

[93] D. J. Ostry and K. G. Munhall, "Control of rate and duration of speech movements," *J. Acoust. Soc. Amer.*, vol. 77, pp. 640–648, 1985.

[94] S. L. Oviatt *et al.*, "A rapid semi-automatic simulation technique for investigating interactive speech and handwriting," in *Proc. Int. Conf. Spoken Lang. Processing* (University of Alberta, Canada), Oct. 1992, vol. II, pp. 1351–1354.

[95] S. L. Oviatt *et al.*, "A simulation-based research strategy for designing complex NL systems," in *Proc. ARPA Human Lang. Technol. Workshop* (Princeton, NJ), Mar. 1993.

[96] D. Pallett *et al.*, "Benchmark tests for the DARPA spoken language program," in *DARPA Workshop Speech, Natural Lang. Processing*, Mar. 1993.

[97] C. L. Paris, "Tailoring object descriptions to a user's level of expertise," *Computational Linguistics*, vol. 3, no. 14, 1988.

[98] A. Pentland and K. Mase, "Lip reading: Automatic visual recognition of spoken words," in *Proc. Image Understanding and Machine Vision, Optical Soc. Amer.*, June 12–14, 1989.

[99] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. IEEE Commun. Soc. Global Telecommun. Conf.* (Atlanta, GA), Nov. 1984, pp. 26–29.

[100] E. D. Petajan, B. Bischoff, and D. Bodoff, "An improved automatic lipreading system to enhance speech recognition," in *Proc. ACM SIGCHI-88*, 1988, pp. 19–25.

[101] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. 1984 IEEE ICASSP*, pp. 18A.2.1–2.4.

[102] P. Price, "Male and female voice source characteristics: Inverse filtering results," *Speech Commun.*, vol. 8, pp. 261–277, 1989.

[103] P. Price *et al.*, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2956–2970, 1991.

[104] R. Reichman, *Getting Computers to Talk Like You and Me*. Cambridge, MA: M.I.T. Press, 1985.

[105] B. Repp, "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," *Psycholog. Bull.*, vol. 92, pp. 81–110, 1982.

[106] D. Roe, F. Pereira, and R. Sproat, "Efficient grammar processing for a spoken language translation system," in *Proc. 1992 IEEE ICASSP*, Mar. 1992, p. I.213.

[107] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Dekker, 1991, pp. 231–268.

[108] R. Schwartz, Y. Chow, and F. Kubala, "Rapid speaker adaption using a probabilistic spectral mapping," in *Proc. 1987 IEEE ICASSP*, pp. 633–636.

[109] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, no. 1, pp. 55–76, 1988.

[110] _____, "A relaxation method for understanding spontaneous utterances," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.

[111] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Commun.*, vol. 5, no. 2, pp. 159–170, June 1986.

[112] E. Shortliffe, *Computer-Based Medical Consultations*. New York: Elsevier, 1976.

[113] E. Shriberg, J. Bear, and J. Dowding, "Automatic detection and correction of repairs in computer-human dialog," in *DARPA Workshop Speech, Natural Lang. Processing*, Feb. 1992.

[114] H. F. Silverman, "Some analysis of microphone arrays for speech data acquisition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 2, pp. 1699–1712, Dec. 1987.

[115] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone-array data," *Comput., Speech, Lang.*, vol. 6, no. 2, pp. 129–152, Apr. 1992.

[116] H. F. Silverman *et al.*,"Experimental results for baseline speech recognition performance using input acquired from a linear microphone array," in *Proc. 5th DARPA Workshop Speech, Natuaral Lang.* (Arden House, Harriman, NY), Feb. 1992.

[117] K. Silverman *et al.*, "TOBI: A standard for labeling english prosody," in *Proc. Int. Conf. Spoken Lang. Syst.* (Banff, Alberta, Canada), Oct. 1992, vol. II, pp. 867–870.

[118] R. W. Smith, D. R. Hipp, and A. W. Biermann, "A dialogue control algorithm and its performance," *3rd Conf. Applied Natural Lang. Processing* (Trento, Italy), Mar. 31–Apr. 3, 1992.

[119] R. M. Stern *et al.*, "Multiple approaches to robust speech recognition," in *Proc. 5th DARPA Workshop Speech, Natural Lang.* (Arden House, Harriman, NY), Feb. 1992.

[120] K. N. Stevens, *Phonetic Linguistics*. New York: Academic, 1985, pp. 243–255.

[121] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *Proc. Int. Joint Conf. Neural Networks*, 1992, pp. II:286–295.

[122] W. V. Summers, "Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analyses," *J. Acoust. Soc. Amer.*, vol. 82, pp. 847–863, 1987.

[123] B. Sundheim, "Overview of the Third Message Understanding Evaluation and Conference," in *Proc. Message Understanding Conf.* (San Mateo, CA), 1991.

[124] P. Suppes, "Current trends in computer assisted instruction," in *Advances in Computers*, M. C. Yovits, Ed. New York: Academic, 1979.

[125] _____, "University-level computer-assisted instruction at Stanford: 1968–1980, " publication of Inst. for Math. Studies in Social Sci., Stanford Univ., Stanford, CA, 1981.

[126] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews, "An investigation of locus equations as a source of relational invariance for stop place categorization," *J. Acoust. Soc. Amer.*, , vol. 90, pp. 1309–1325, 1991.

[127] W. R. Swartout, "XPLAIN: A system for creating and explaining expert consulting systems," *Artif. Intell.*, vol. 3, no. 2, pp. 285–325, 1983.

[128] H. Tennant, "Experience with the evaluation of natural language question answerers," Tech. Rep., Univ. of Illinois, Urbana-Champaign, 1979.

[129] H. Traunmuller, "Perceptual dimension of openness in vowels," *J. Acoust Soc. Amer.*, vol. 69, pp. 1465–1475, 1981.

[130] D. V. Bergem, L. Pols, and F. Koopmans van Beinum, "Perceptual normalization of the vowels of a man and a child," *Speech Commun.*, vol. 7, pp. 1–20, 1988.

[131] A. Varga et al., "Noise compensation algorithms for use with hidden Markov model based speech recognition," in Proc. 1988 IEEE ICASSP.

[132] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in Proc. 1990 IEEE ICASSP, pp. 845–848.

[133] E. Vatikiotis-Bateson and J. A. S. Kelso, "Rhythm type and articulatory dynamics in English, French, and Japanese," J. Phonetics, vol. 21, 1992.

[134] N. Veilleux and M. Ostendorf, "Prosody/parse scoring in ATIS," in Proc. Workshop Human Lang. Technol., Mar. 1993.

[135] A. Waibel et al., "JANUS: A speech-to-speech translation system using connectionist and symbolic processing strategies," in IEEE ICASSP, May 1991.

[136] D. E. Walker, Ed., Understanding Spoken Language. New York: North Holland/Elsevier, 1978.

[137] M. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," unpublished, 1992.

[138] W. Ward and S. Young, "Flexible use of semantic constraints in speech recognition," in Proc. 1993 IEEE ICASSP (Minneapolis, MN), Apr. 1993, pp. II:49–50.

[139] C. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases," in Proc. IEEE ICASSP, May 1991, pp. 321–324.

[140] _____, "Automatic recognition of intonational features," in Proc. IEEE ICASSP, Mar. 1992.

[141] M. Witbrock and P. Haffner, "Rapid connectionist speaker adaptation," in Proc. 1992 IEEE ICASSP, Mar. 1992, pp. I:453–456.

[142] U. Wolz, K. R. McKeown, and G. Kaiser, "Automated tutoring in interactive environments: A task centered approach," J. Machine Mediated Learning, 1989.

[143] S. R. Young et al., "High level knowledge sources in usable speech recognition systems," Commun. ACM, vol. 32, no. 12, pp. 183–194, Feb. 1989.

[144] B. P. Yuhas, Jr. M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," IEEE Commun. Mag., Nov. 1989.

[145] B. P. Yuhas, Jr. et al., "Neural network models of sensory integration for improved vowel recognition," Proc. IEEE, vol. 78, no. 10, pp. 1658–1668, 1988.

[146] V. W. Zue, "The use of speech knowledge in automatic speech recognition," in Proc. IEEE, pp. 1602–1615, 1985.

Ron Cole, photograph and biography not available at time of publication.

Lynette Hirschman, photograph and biography not available at time of publication.

Les Atlas, photograph and biography not available at time of publication.

Mary Beckman, photograph and biography not available at time of publication.

Alan Biermann, photograph and biography not available at time of publication.

Marcia Bush, photograph and biography not available at time of publication.

Mark Clements, photograph and biography not available at time of publication.

Jordan Cohen, photograph and biography not available at time of publication.

Oscar Garcia, photograph and biography not available at time of publication.

Brian Hanson, photograph and biography not available at time of publication.

Hynek Hermansky, photograph and biography not available at time of publication.

Steve Levinson, photograph and biography not available at time of publication.

Kathy McKeown, photograph and biography not available at time of publication.

Nelson Morgan, photograph and biography not available at time of publication.

David G. Novick,, photograph and biography not available at time of publication.

Mari Ostendorf, photograph and biography not available at time of publication.

Sharon Oviatt, photograph and biography not available at time of publication.

Patti Price, photograph and biography not available at time of publication.

Harvey Silverman, photograph and biography not available at time of publication.

Judy Spitz, photograph and biography not available at time of publication.

Alex Waibel, photograph and biography not available at time of publication.

Clifford Weinstein, photograph and biography not available at time of publication.

Steve Zahorian, photograph and biography not available at time of publication.

Victor Zue, photograph and biography not available at time of publication.