Workshop on Spoken Language Understanding[1]

A Workshop Sponsored by the National Science Foundation

September 1, 1992

Ron Cole
Oregon Graduate Institute

Lynette Hirschman
MIT

Les Atlas
Univ. of Washington

Mary Beckman
Ohio State Univ.

Alan Biermann
Duke Univ.

Marcia Bush
Xerox Palo Alto Research

Mark Clements
Georgia Inst. of Technology

Jordan Cohen
IDA Center for Comm. Research

Oscar Garcia
George Washington Univ.

Brian Hanson
Speech Technology Laboratory

Hynek Hermansky
U S WEST

Steve Levinson
AT&T Bell Labs

Kathy McKeown
Columbia Univ.

Nelson Morgan
ICSI, UC Berkeley

David Novick
Oregon Graduate Inst.

Mari Ostendorf
Boston Univ.

Sharon Oviatt
SRI International

Patti Price
SRI International

Harvey Silverman
Brown Univ.

Judy Spitz
NYNEX AI Speech Group

Alex Waibel
Carnegie-Mellon Univ.

Cliff Weinstein
MIT Lincoln Laboratory

Steve Zahorain
Old Dominion Univ.

Victor Zue
MIT

[1] Suggested citation: R.A. Cole and L. Hirschman et al., Workshop on Spoken Language Understanding, Oregon Graduate Institute Technical Report No. CS/E 92-014, Sep. 1, 1992.

Preface

The February, 1992 NSF Workshop on Spoken Language Understanding brought together scientists from a number of disciplines to identify research directions needed to produce spoken language systems. This report describes the key research topics, the expected benefits of the research, and recommendations to NSF on the infrastructure needed to support the research.

The Workshop was supported by Grant No. IRI-9208831 from NSF awarded to Ron Cole of the Oregon Graduate Institute, Lynette Hirschman of the Massachusetts Institute of Technology, and Steve Zahorian of Old Dominion Univeristy. The workshop organizing committee was Ron Cole, Lynette Hirschman, Alex Waibel, Steve Zahorian and Victor Zue. The workshop report was put together by Ron Cole and Lynette Hirschman.

The individual sections were authored by the following: **1 Executive Summary:** Lynette Hirschman; **2 Introduction:** Ron Cole and Lynette Hirschman; **3 Research Directions:** 3.1 Robust Speech Recognition: Steve Zahorian, with help from Mary Beckman, Mark Clements, Brian Hanson, Hynek Hermansky, Nelson Morgan and Harvey Silverman; 3.2 Automatic Training and Adaptation: Lynette Hirschman; 3.3 Spontaneous Speech: David Novick with help from Patti Price, Mari Ostendorf and Lynette Hirschman; 3.4 Dialogue Models: Alan Bierman, with help from Lynette Hirschman and Kathy McKeown; 3.5 Natural Language Response Generation: Kathy McKeown with help from Mari Ostendorf; 3.6 Speech Synthesis and Generation: Mari Ostendorf, with help from Patti Price; 3.7 Multilingual Systems: Cliff Weinstein and Steve Levinson; 3.8 Interactive Multimodal Systems: Sharon Oviatt, Marcia Bush and Ron Cole; **4 Infrastructure:** 4.1 Multi-disciplinary Research and Training: Victor Zue; 4.2 Post-doctoral funding: Les Atlas; 4.3 Database Development and Sharing: Jordan Cohen; 4.4 Computational Resources: Nelson Morgan and Steve Levinson; 4.5 Sharing of Speech Research Tools and Algorithms: Ron Cole; 4.6 Communication: Alex Waibel; 4.7 US Science Institute (NSF Initiative): Jordan Cohen; **5 Benefits:** 5.1 Student Education and Jobs: Mari Ostendorf and Les Atlas; 5.2 U.S. Competitiveness in the Global Marketplace: Judy Spitz; 5.3 International Cooperation and Business: Steve Levinson and Alex Waibel; 5.4 Societal Impact of Spoken Language Systems: Steve Levinson and Oscar Garcia; 5.5 Benefit to Scientific Community: Les Atlas; **6 Recommendations:** Ron Cole and Lynette Hirschman.

Contents

1 Executive Summary 1

1.1 The Grand Challenge 1

1.2 Key Research Areas 1

1.3 Infrastructure 2

1.4 Recommendations 2

2 Introduction 4

2.1 Importance of the Problem 5

2.2 Worldwide Funding Profile 5

2.3 NSF's DARPA dilemma 6

2.4 Scope and Organization of the Workshop 7

2.5 Organization of the Report 8

3 Research Directions 9

3.1 Robust Speech Recognition 9

3.2 Automatic Training and Adaptation 15

3.3 Spontaneous Speech 18

3.4 Dialogue Models 21

3.5 Natural Language Response Generation 25

3.6 Speech Synthesis and Generation 29

3.7 Multi-lingual Systems 32

3.8 Interactive Multimodal Systems 34

4 Infrastructure 39

4.1 Multi-disciplinary Research and Training 39

4.2 Post-doctoral funding 40

4.3 Database Development and Sharing 41

4.4 Computational Resources 41

4.5 Sharing of Speech Research Tools and Algorithms 44

4.6 Communication 44

4.7 US Science Institute (NSF Initiative) 46

5 Benefits **47**

5.1 Student Education and Jobs 47

5.2 U.S. Competitiveness in the Global Marketplace 47

5.3 International Cooperation and Business 48

5.4 Societal Impact of Spoken Language Systems 49

5.5 Benefit to Scientific Community 50

6 Recommendations **52**

1 Executive Summary

1.1 The Grand Challenge

A spoken language system integrates speech recognition (to identify the words), natural language processing (to understand what the words mean) and interface technology (to provide the appropriate response). Current prototype spoken language systems now provide near real-time performance on small (1000-word) tasks, such as interactive air travel planning or urban navigation. Such systems will change the paradigm of human-computer interaction from "programming" to "conversation", and radically expand access to on-line resources, particularly for novice users, handicapped users, and people with hands-busy applications.

The NSF has a crucial role to play in maintaining the US technological lead in this area, in the face of well-funded efforts in Europe and Japan. NSF can complement the current DARPA program by continuing its support of education and training, by supporting high-risk, innovative efforts, and by building an infrastructure of common hardware and software resources that will encourage data and resource sharing.

1.2 Key Research Areas

The report identifies key research areas requiring support. These are:

- **Robust Speech Recognition** – so that systems degrade gracefully when information is lost due to limited bandwidth, background noise, channel distortion, etc.
- **Automatic Training and Adaptation** – to make systems easy and cheap to adapt or train for 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 a user.
- **Natural Language Response Generation** – so that the system provides coherent, appropriate communication with the user.
- **Speech Synthesis and Generation** – to produce comprehensible output to the user and to enhance our understanding of speech.

- **Multi-lingual Systems** – to provide speech to speech language translation and support information access in a multi-lingual society and a global economy.

- **Multi-modal 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.

1.3 Infrastructure

NSF plays a critical role in the educational and computational infrastructure for spoken language research. This includes support for:

- Training of new scientists at the doctoral and post-doctoral level;
- Developing new courses and continued cross-training of researchers;
- Creating corpora and evaluation methods to develop systems and measure their progress;
- Providing researchers a basic configuration of workstations, plus appropriate storage, accompanied by a common set of software tools;
- Communicating research results, data, and ideas across sites; and
- Creating a science institute at a national level to create, co-ordinate and support the educational and computational infrastructure, for this and other "Grand Challenge" research areas.

1.4 Recommendations

The workshop makes the following five recommendations to NSF:

1. NSF funding for spoken language research needs to be significantly higher to overcome fundamental scientific difficulties in a discipline with well-known, significant payoffs for society.
2. NSF funding should focus on basic research and new approaches to spoken language, as outlined in this report.
3. NSF should encourage collaborative and interdisciplinary research in spoken language understanding.

4. NSF should provide infrastructure support, including educational support, common resources, and support for communication among scientists.

5. NSF should fund a broad spectrum of research, including basic research, system development, and data collection.

2 Introduction

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, phone 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.

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 multi-lingual interactions.

Spoken language system technology has made rapid advances in the past decade, supported by progress in the underlying technologies. 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 300–1000 words, and an overall understanding rate of about 70% [91, 22, 108, 3, 24].

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. First, 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. Second, systems 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. Third, systems must not only be real-time and large enough to handle "real" applications, but they must be easy and cheap to build or adapt for new applications. Finally the hardware supporting them must be cheap and "built-in", so that spoken language interfaces become as natural as a keyboard or mouse is used today. These systems will eventually become part of multi-model systems, in which the user's intent is derived by combining speech with facial expressions, eye movements, gestures, handwriting, and other input features, and in which the machine communicates with the user through multi-media responses. In addition, future systems will be multi-lingual, performing speech-to-application translation or even speech-to-speech translation.

2.1 Importance of the Problem

Spoken language has been designated as one of the "Grand Challenge" applications for the NSF High Performance Computing and Communication (HPCC) Program. This technology is of critical national importance because of its:

- Potential to change how people interact with computers from a "programming" model to a "conversational" model by providing a convenient and natural modality (speech) to access and manipulate on-line information;
- Potential to make on-line information resources readily available to vast new classes of users (casual users, novices, handicapped users);
- Ability to position US industry to capitalize on its research leadership in spoken language, speech, natural language, and human-machine interface technologies;
- Ability to support international cooperation, diplomacy, and commerce in the increasingly interconnected global economy via multi-lingual and speech-to-speech translation systems.

2.2 Worldwide Funding Profile

Human machine interaction, with emphasis on spoken language understanding, is now a national priority in the United States, Japan, Korea, and the European Community (particularly Britain, Germany, France, and Italy). Each of these countries has major programs, funded at the level of $20 to $30 million per year, to advance the state of the art of spoken language systems. These projects include the ATR research laboratory for automatic speech translation in Japan, the ESPRIT program in Europe, and the DARPA speech and natural language program and the DARPA neural net program in the United States. The U. S. government also sponsors research through Rome Air Development Center, AFOSR, ONR, and through funding of spoken language research at government laboratories.

In addition to these major programs, there is substantial corporate investment in research and development of speech and natural language technology. In

Japan, NTT and NEC have major efforts, while in Europe, the major efforts are at Siemens and Phillips. In the United States, IBM has maintained a 10 to 20 person speech recognition group continuously over the past fifteen years. AT&T research laboratories invests about $30 million per year in speech research. A number of other companies maintain substantial speech efforts, including Apple Computer and NYNEX. In addition to these major efforts, there are dozens of companies that maintain smaller efforts of three to ten researchers. The annual cost of research in speech recognition in the United States is easily over $200 million. However, while industry provides substantial funding for speech and spoken language research, most of the results are proprietary or company specific. There is a need for publicly available results and systems that industry-funded programs cannot supply.

The total budget for NSF for speech and natural language research is $2–3 million. This is a remarkably small number, considering that NSF is the main source of basic, non-proprietary research in the United States, and the main source of funding focused on non-military applications. We estimate that NSF funding of speech and natural language is less than 1% of the total funding profile in the United States, and no more than 5% of the total government funding in this area.

2.3 NSF's DARPA dilemma

The success of the DARPA speech and natural language program produces a dilemma for NSF: how can NSF, which funds small individual research grants, hope to make a significant impact on science and technology in light of the DARPA speech and natural language program? The DARPA program has significantly more funding (by a factor of 5 to 10), involvement of the major research laboratories (e.g., CMU, MIT, BBN, SRI, MIT Lincoln Labs), the advantage of continuity, since the same sites are funded each year, a strong infrastructure consisting of database collection, development and distribution, a rigorous evaluation methodology, a rich history, and high visibility.

The answer to this question can be found in the nature of the two programs. The DARPA program is designed to promote the development of spoken language systems using current technology. The program is task-oriented, with progress measured in terms of the performance of the resulting systems. Moreover, the program is structured as a competition, so each site is evaluated annually on each new task using a common set of training and test data. While this approach has produced steady progress and advances in the capabilities of spoken language systems, the role (and time for) basic research is minimized. There is pressure for DARPA sites to minimize risk-taking in favor of steady incremental progress, so that groups tend to coalesce around a proven approach, rather than exploring riskier alternatives (for example, six of the seven major speech groups now use the Hidden Markov Model approach).

NSF, therefore, has a major role to play in the field of human machine interaction by funding basic research, encouraging new and high risk approaches, and supporting educational and training activities, to ensure a steady flow of well-trained scientists to tackle these problems. This will provide the theoretical foundation and research breakthroughs to keep US technology at the forefront of this rapidly developing field. The workshop participants agreed that NSF and DARPA should play complementary roles: NSF can leverage the advances and infrastructure of the DARPA program to support basic research at multiple sites, including a number of single-PI sites (in contrast to the 5–15 person groups supported by DARPA). NSF should focus on innovative approaches and component-level advances, and rely on the DARPA infrastructure, with its emphasis on system-building and system-level evaluation, to integrate these technologies.

2.4 Scope and Organization of the Workshop

This report is the result of a workshop on spoken language understanding held in Washington D.C. on Feb 10th and 11th, 1992. The workshop was organized at NSF's request to help identify areas of research that deserve future funding, to describe the infrastructure needed to allow researchers to make significant progress, and to recommend funding strategies that will leverage NSF's unique role in supporting basic research and training of future scientists.

The original topic of the workshop was spoken language understanding, with emphasis on speech and natural language understanding. It was soon recognized that issues in speech recognition and natural language understanding are intimately related to issues of speech production, response generation and dialogue modeling, so the scope of the workshop was expanded to spoken language systems, including multi-lingual, multi-modal and multi-media systems.

The specific goals of the workshop were:

1. To identify the most important areas of research in speech and natural language understanding, and, in particular, to identify those not currently addressed by NSF or other funding agencies;
2. to determine how NSF could benefit from other programs, such as the DARPA speech and natural language program, while avoiding competition and duplication of effort;

3. to produce a set of recommendations to NSF to help guide funding opportunities, and provide the necessary educational and research infrastructure.

The workshop was run as a set of coordinated plenary sessions and "breakout sessions" consisting of working groups on four general topic areas: robust systems, integration of speech and natural language, human-machine interaction, and multilingual systems.

- **Session 1– Introductions:** The workshop began with program reviews by Y. T. Chien and John Hestenes of NSF, and Charles Wayne of the DARPA speech and natural language program. This was followed by three minute presentations by each participant describing their area of interest and thoughts about future research priorities.
- **Session 2 – Research Issues:** The second session consisted of meetings of the four breakout groups. Each group generated a list of important research topics, which were then discussed in the following plenary session by all participants.
- **Session 3 – Research Goals:** In the third session, the breakout groups met to determine a prioritized list of research goals, and to produce recommendations about the infrastructure needed to achieve these goals. These were then modified during the following plenary session.
- **Session 4 – Conclusions:** During the fourth and final session, the workshop report was outlined and sections were assigned to authors.

2.5 Organization of the Report

The remaining sections are organized as follows: Section 3 describes the major areas of research needed to produce spoken language systems; Section 4 describes the infrastructure needed to support the research; Section 5 describes the expected benefits of the research; and Section 6 summarizes the recommendations of the workshop.

3 Research Directions

We identify eight areas of research that are critical to progress in the development of spoken language systems. In each area, we identify the nature of the problem, the key research challenges and fundamental scientific issues, and the benefits of the research.

3.1 Robust Speech Recognition

Robustness in speech recognition can be defined as minimal, graceful degradation in performance with changes in input conditions, such as additive noise, different speakers, or other small (insofar as human listeners are concerned) systematic changes in the acoustic signal. At present speech recognition systems are notably non-robust, with performance degrading significantly with modifications as minor as a change in microphone. Because of this, systems trained in the laboratory usually fail when exposed to operating conditions in the field. Although many signal processing strategies offer partial solutions [6, 84, 4, 19, 58, 72, 50, 45], the robustness problem is far from solved.

The fundamental method for improving robustness is to better understand the many sources of variability in the speech signal, and to develop features and algorithms which are sensitive to the variability of interest and relatively less sensitive to other sources of variability. To accomplish this objective, all pertinent sources of variability must be understood, modeled, and properly accounted for.

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 channel to the machine.

Recognition technology has matured to the point where it is no longer either necessary or acceptable to systematically ignore many of the variabilities shown in the figure. For example in early systems, only close-talking microphones in noise free rooms were used, to avoid many of the variabilities due to acoustical factors. Specifically-chosen talkers were used to avoid explicitly modeling the talker's physical characteristics and dialect. Difficult channels (e.g. long line telephone) were only simulated. Over-simplified models that average over segmental and prosodic contexts were used in order to avoid these sources of variability. Today, however, every one of the sources of variability can and should be addressed: modeling of variability constitutes an important focus for study.

Over the last few years, performance of speech recognition systems have been improved largely through the use of statistical modeling techniques which account

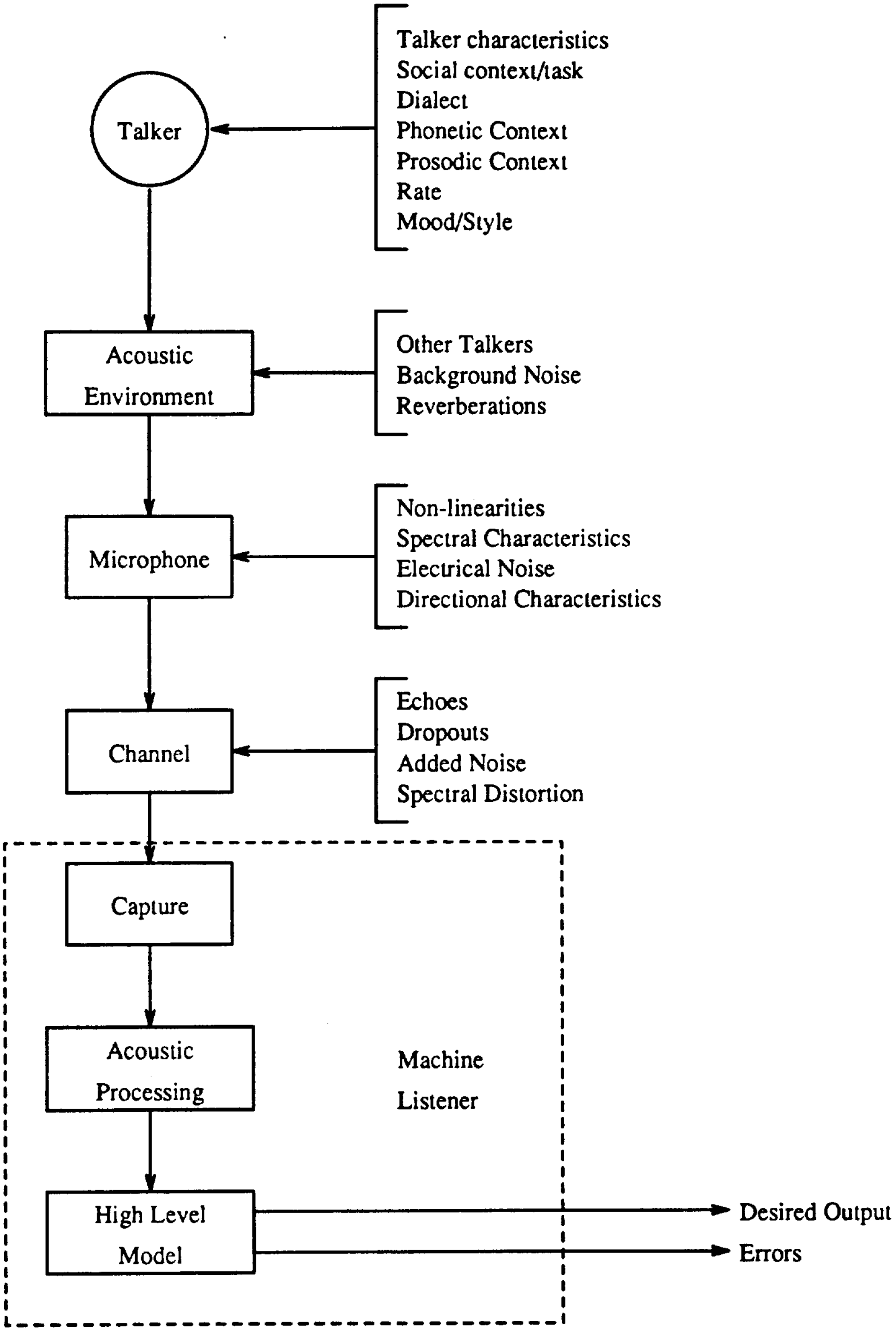

Figure 1: Sources of variability in the speech signal (from the viewpoint of an automated speech recognition system).

for some of the variability. However, in all but highly constrained tasks, automated speech recognition performance is unacceptably low. Although statistical techniques do play a valuable role in recognition, it is imperative to address these sources of variability directly by developing better models and new representations using natural speech spoken in real environments.

There is a continuing need for new ideas, new models, and unusual approaches to address the problem of robustness. 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 Figure 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 [14].

More robust speech recognition will be achieved by more explicit modeling of the many sources of variability in the speech signal and by treating some of these variabilities as knowledge sources rather than noise sources. Much of the variability in the acoustic properties of particular speech sounds arises through systematic modifications in timing and rates of movement of the articulators used to produce the sounds. Search for the underlying causes of variability in the speech signal, therefore, must include study of the influence of perturbations in the production of this signal. Modeling of sources of variability will require algorithm development, computer resources, and large amounts of experimental testing. Many of the relevant issues can be addressed using the current NSF funding model, given appropriate communication among sites and shared resources and cooperation. Because many of these new ideas can be best tested as additions or modifications to existing systems, it is absolutely essential that an extremely well-developed, easily usable set of tools be available to principal investigators.

Finally, we need a better understanding of the language production process in the context of natural dialogues to model the higher-level sources of variability needed to produce robust systems. The sections occuring later in this report on spontaneous speech, dialogue modeling, and response generation deal with these issues directly.

3.1.1 Key Research Challenges and Fundamental Scientific Issues

1. **Developing Better Models of the Talker and the Task.**

   Robust systems will result from a deeper understanding of the sources of variability in speech. As this understanding emerges, what looks like noise at one level can provide useful information for interpreting the message at another level. The spectral characteristics of a sound segment vary tremendously from one linguistic context to another (e.g., [98, 67, 117, 70, 88, 25, 139, 104]; see [103] for a partial overview). At present, coarticulatory variability due to segmental context is often accounted for with context-dependent Hidden Markov Models. More explicit models could improve on the large gains achieved with the context-dependent HMM's. This is particularly the case with the coarticulatory variability that is conditioned by prosodic structure. For example, in American English, the phoneme /t/ is released with aspiration at the beginning of a stressed syllable, is realized as a glottal stop in syllable final position, and is realized as a short voiced flap intervocalically between a stressed and unstressed syllable (see, [33, 46, 32, 106, 36, 10, 97] for other examples). In current approaches, this variability complicates recognition by vastly increasing the number of contexts with which a recognizer must be trained. However, if the recognition system incorporates a representation of these prosodic patterns and a model of the ways in which stress and syllable structure condition segmental variability, the signal can be parsed into stressed and unstressed syllables (see, e.g., [38, 13, 101, 96]), thus vastly reducing the lexical search space. Furthermore, with appropriate transformation of the signal to emphasize the common articulatory aspects of the sound, the apparent variability in the signal can be greatly reduced.

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., [34, 89, 129]), local changes in speaking rate might be used to recognize such prosodic patterns as stress or phrase-final lengthening (e.g., [120, 20, 100]); more global changes might help parse changes in topic or conversational turn (see e.g., [8, 7, 53]) and even some more intricate pragmatic differences among utterances ([54]). Modeling this "noise" at the phonetic level thus may add one more type of information to the syntactic, semantic, and pragmatic constraints that are used to understand the signal at the higher level of discourse understanding.

Speech recognition models could also benefit from interdisciplinary approaches that incorporate models accounting for dialect differences and social context effects on the discourse (e.g., [64, 42, 41, 49, 66, 56, 11, 18]). By incorporating explicit representations of phonological differences among dialects, for example, a recognition system could vastly reduce the amount of training data necessary to adapt to a new set of talkers from a different dialect specification. If the dialect is an "r-drop" variety that distinguishes between words such as "mark" and "mock" by vowel quality rather than by presence versus absence of a post-vocalic /r/ consonant [65], the recognizer might be trained to locate this distinction correctly in the vowel and not look for the spectral correlates of the /r/. Analogous adjustments to representations can be made at the lexical, morphological, and syntactic level. Explicit modeling of such dialect variation at the phonetic level will be particularly important if it turns out, as suggested by [87], that patterns of coarticulatory variation across word boundaries can differ from one dialect to another.

As a final example, consider the talker's personal voice characteristics. Until quite recently, speaker-independent recognition was accomplished by training on large amounts of data from many different talkers. Some systems have begun to model specific talker characteristics in a more explicit way, using adaptation [44, 133, 105]. There is a great deal of basic research data on talker characteristics and speaker normalization from fields as diverse as speech physiology, phonetics, psychoacoustics, and speech synthesis (e.g., [12, 79, 127, 31, 86, 43, 128, 85, 125, 99, 57, 61, 60], etc.). Incorporating these results into recognition systems could lead to more illuminating models and representations, with implications not only for speech recognition but also for talker verification/identification, as well as maintaining voice characteristics for speech synthesis (e.g., for a telephone translation system).

2. **Acoustic environment and microphone.**

   It is widely accepted that computer users do not want to be encumbered by a headmounted or hand held microphone to converse with a computer. Thus a remote microphone attached to the system, which can track a talker and maintain consistently high quality is highly desirable. There is currently some effort to understand how to adapt a system trained on one microphone to perform at full capability using a different microphone/environment [48, 50, 116]. Multiple microphone systems also offer promise, with the prospect of tracking a remote talker with high quality [29, 111, 15, 5, 30, 112, 113]. While there has been some success with these methods, considerable work is still necessary. 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.

3. **Channel.**

   Speech recognition over the telephone line is imperative for many commercial applications of human-computer interface technology. For many years people have been addressing problems of echoes, noise, nonlinearities, and spectral distortions of the telephone channel. Some preliminary work in this area suggests that fairly simple engineering approaches to filtering out these effects may be fairly effective. However, more comprehensive investigations are required.

4. **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. Thus, speech components emphasized by a conventional automated speech recognition front end might be misleading for automatic decoding of linguistic messages.

   Some recently developed speech analysis techniques attempt to model the basic properties of human speech perception [107, 35, 74, 47]. Such approaches are reported to significantly improve the robustness of automated speech recognition with respect to improved handling of non-linguistic sources of variability in speech such as differences due to talkers, differences in the acoustic environment, or overall spectral differences due to a change of microphone or microphone position [48, 50, 14]. Thus, it appears that even rather simple perceptual models of human speech perception hold promise for alleviating many of the problems in automatic speech recognizers. 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., [69, 123, 118]). Most large commercial speech research centers tend to shy away from non-engineering disciplines. Therefore, NSF could try to fill this important knowledge gap by encouraging the study of human speech perception, which would also serve to attract industrial attention and support to examine the practical implications of this knowledge.

3.1.2 Benefits of Robust Systems

The bottom line in automatic speech recognition technology is that it does not work well enough to be reliably used in most real world applications [138]. Users will naturally be reluctant to rely on automatic speech recognition as a computer input device if they have to talk in a machine-like way, if it must be highly tuned to their particular voice, if it fails on a day when they have a cold, if performance drops severely when there is extra background noise, or if it generally does not perform well. However, as speech recognition technology becomes more robust, with far fewer errors and much more graceful degradation, the number of potential applications is extremely large.

3.2 Automatic Training and Adaptation

One of the major obstacles to building and evaluating spoken language systems is the high cost (in terms of labor) of porting a system to a new application. This is true for both speech recognition and natural language components of a system. 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 DARPA Air Travel (ATIS) domain, joint data collection activity across five sites resulted in the collection of over 15,000 utterances [23]. The data collection activity represents several person months of effort at each site, not counting the costs of checking and distributing the data.

Most natural language understanding systems require training data too, but 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 natural language processing experts to do the development. In addition, evaluation of natural language systems is still highly labor-intensive. It requires the specification of a "correct *answer*", *and annotation of* both training and test data to provide those answers. The annotation of 5000 training sentences for the ATIS task required more than twelve months of effort.

Multi-lingual systems represent a significant challenge for portability as well. Porting a system to a new language often requires a complete rewrite of all components of the system, in addition to the need for new training data for both speech and language.

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.

3.2.1 Key Research Challenges And Fundamental Scientific Issues

1. **Better Use of Training Data.**

   If we require a whole new set of training data for each application, portability will remain an expensive undertaking. There has been work recently on task-independent vocabulary modeling [55]. These notions also need to 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.

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 from other talkers, breath noises, coughs, filled pauses, and environmental noises such as phone 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. Words need to be marked for part of speech (and other syntactic information, such as complement structure) if parsing is involved. In addition, words need to be marked for semantic class, and for 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. Research in part-of-speech tagging and discovery of syntactic and semantic classes is necessary to automate portability between tasks and also between languages, for multi-lingual and translation applications.

4. **Automated Training for Natural Language Systems.**

   The natural language community has been influenced by the successful use of stochastic models in the speech community. It is important to support continued research in this area, particularly on methods of combining the knowledge-based paradigms in use for natural language and AI systems with the stochastic models used in speech recognition systems.

5. **Shared Lexicon.**

   We should find ways to utilize existing repositories of "expert information". Large lexicons are a good example, and under the Consortium for Lexical Research, these resources are being made increasingly available. However, we need better tools to extract knowledge from these resources, we need to understand how to combine information from multiple sources (lexicons), and how to best incorporate such information to build robust and portable spoken language systems.

6. **Knowledge Engineering Bottleneck.**

   One of the most labor and expertise intensive tasks is the construction of a domain model, providing a taxonomy of the objects in the domain, and their relations to each other. Related to this is the problem of linking 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 ways to automate this, particularly in the context of building a spoken language system.

7. **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.

8. **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 [121] 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.

3.2.2 Benefits of Research on Automatic Training and Adaptation

The introduction of spoken language systems into a variety of real world applications requires fundamental research on how to quickly adapt these systems to new applications. If this area of research is ignored, the possibility exists that our competitive advantage in developing spoken language systems technology will be lost to those who are more readily able to apply it to new applications. Thus, research in system development must proceed in parallel with research on rapid training and adaptation.

Spoken language systems bring together speech recognition and natural language understanding technology. A key aspect of this "coming together" has been cross-fertilization, bringing renewed interest in the use of statistical modeling applied to language phenomena, integrated with a priori knowledge sources, such as syntactic classes, or even "kernels" of semantic classes. This is an extremely important area of research; success here will make it possible to port systems quickly and cheaply to new applications and languages.

3.3 Spontaneous Speech

The field of spoken language understanding combines research in speech recognition and natural language processing. But for both speech recognition and natural language processing, our technology has largely been based on the written—as opposed to the spoken—form of language. Speech recognition research and evaluation has focussed on read sentences; natural language research has focussed on written language. Text-based models, however, are not adequate for understanding spontaneous language because there are significant differences between written and spoken language. Spontaneous speech is more variable than text, which has conventions imposed by its written form. In fact, spontaneous speech is notoriously ill-formed, full of "improper" usages, mismatched agreements, 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 non-speech noises (inhalation, cough); (2) meta-sentential events such as correction and editing ("Denver, I mean BOSTON"), and (3) non-verbal communication (eye contact, nodding) that play a part in maintaining a conversation.

Though the prior research focus on read speech and written text has meant that our knowledge of spontaneous speech is limited, we know that there are regularities associated with spontaneous speech phenomena. For example, repairs occur with some frequency (around 6 percent of sentences in rather planned spontaneous data such as ATIS [110], and in 34 percent of sentences in a human-human dialogue corpus [68]), 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. There are three major reasons for the failure of current models: first, repairs may fall outside the language model; second, most language understanding systems do not have the ability to "overwrite" the mistake with its repair to form a repaired semantics; and third, the prosodic cues that help humans detect these events are often ignored by the current generation of systems. By studying these phenomena and by explicitly incorporating them in both the language model (for interpretation) and the acoustic model (for detection), we should be able to build much more robust spoken language systems.

Prosody is an important component of spoken language that is not well represented in written language, hence it has often been ignored in speech recognition and in language understanding. However, prosody may provide the "glue" and the "pointers" that make spoken language so 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 [62].

Another issue that deserves further investigation is turn-taking, namely how conversational participants signal that they want to take the floor, that they are ready to release the floor, that they require clarification or that they have understood. Research on turn-taking dynamics and conversational control will contribute to our understanding of human-human interaction, and also will make important contributions to building better, more natural and usable human-machine interfaces.

3.3.1 Key Research Challenges and Fundamental Scientific Issues

Spontaneous speech phenomena can be viewed not simply as "linguistic chaff" to be discarded, but rather as 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 consequences beyond linguistic interaction, with significant implications for the design and development of human-computer interfaces. That is, spontaneous speech phenomena may indicate underlying principles of communicative interaction.

Accordingly, the NSF research program in spontaneous speech should stress the following issues:

1. **Computational models of spontaneous speech.**

   There has been relatively little work on spontaneous speech, and almost no work on computational models of spontaneous speech. Such models are necessary at all levels – acoustic, linguistic and prosodic – in order to detect and correct 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. **Understanding of spontaneous speech phenomena.**

   Though it will be an important advance simply to recover from or ignore a spontaneous speech event, we would also like to understand the information (or meta-information) that these phenomena convey about the speaker and the dialogue. We need to develop plausible theories of speech that can account for the range of observed behaviors within an integrated framework.

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 related 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 TI [37]).

3.3.2 Benefits of Spontaneous Speech

The ability to deal with spontaneous speech phenomena is an important property of robust systems. Systems that can not 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 liberate users of spoken language systems from static, stilted interfaces by enabling more natural dialogue interaction.

3.4 Dialogue Models

Dialogue processing is the enabling technology for speech recognition systems. While speech recognition technology may provide better and better guesses at what word tokens were uttered, the machines will not properly use these guesses unless they acquire the user's meaning from those tokens and efficiently respond to the user's desire. The impressive recognizers developed in recent years will remain laboratory curiosities until software systems come into existence that can use them to deliver function to the user.

True speech understanding requires that individual utterance meanings be understood in the context of the larger dialogue structure [1, 40]. This structure must co-ordinate 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 global goal-subgoal 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. From the utterance, knowledge is attached to the global data structures resulting in a net augmentation of those structures representing the specifics of a particular user-system interaction. From the global structures, however, it is often the case that information must be passed down 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 utterance level with information from the discourse level.

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.

3.4.1 Key Research Challenges And Fundamental Scientific Issues

We identify six central areas for research in this field:

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, 71, 102]. The individual subdialogues provide what is called "focus" by [39], and the tracking of subdialogues is called "plan recognition" by [2]. 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 [52]. 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 Bayesian estimates at the signal processing level for improved recognition [21, 16].

   This leads to a new formulation of the speech recognition 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 recognition could reduce perplexity and provide improved error correction. See, for example, the contribution by [135].

   The contribution of the dialogue model to overall system robustness can be dramatic for a variety of reasons. First, speech recognition can be enhanced as already noted. Second, the structure of the interaction enforced by the model imposes a coherence on the long term interaction that will persistently seek success even in the face of erratic behavior at the sentence by sentence level. Thus composite system performance can exceed the quality achieved at atomic levels and deliver increased overall robustness.

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 [114]. 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 [28, 63, 92]. 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 iteractively 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. 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 [77, 75, 82]. 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 subsections on response generation and speech synthesis, below.

3.4.2 Benefits

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. The user need input only short fragmentary utterances that will 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 must necessarily be obtained before speech systems will come into common use.

A generation of researchers interested in these areas is needed to implement this program of research. This field has been underfunded and generally ignored up to this time. The lack of acceptance of speech recognition in the world at large is a direct result of the lack of attention given to the environments in which it might be used.

3.5 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. 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 80's which showed why such an approach is inadequate [59, 51, 78, 81, 83].

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, questions requesting the definition of terms or questions about the differences between concepts [76, 126]. 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 [124] that a simple "translation" of the underlying inference trace (as is often done using templates [109]), 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.

3.5.1 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. This 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 as 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, 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 theses sources be shared? For example, in order to provide responses that are sensitive to user and 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 have 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 what is 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 when a generation system is successful. 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, among others. Each one of these has potential problems. For example, task completion measures definitely 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 that 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 overly rely 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.

3.5.2 Benefits of Research on Response Generation

Response generation is needed for spoken language systems to communicate with the user. This is particularly true of an audio-only medium like the telephone where there is no possibility for using graphical or tabular responses. Although it is possible to convey certain kinds of information without response generation in systems with other channels, this technology is clearly integral in building dialogue-based systems that can support users in complex problem solving and information access activities.

Response generation is also required for machine translation, both written and spoken; without it, there is no way to produce the final translation.

Help system interfaces (particularly for distributed programming environments), computer aided instruction, and task instruction provide other clear examples where traditional approaches (canned text, key word retrieval) are inadequate. In fact, spoken language interfaces have not often been attempted for these applications. Typical help systems provide much more information than is needed to solve the problem at hand and often make it difficult to find the bit of information needed to complete a task [134]. Response generation would allow for a concise answer addressing user problems.

Finally, while many systems are primarily passive and let the user guide the interaction by asking questions, if we are to allow the system to take a more active role, guiding the user to the solution needed by asking appropriate questions, again response generation is needed. A system which can both guide and answer questions would allow for more natural human-computer interaction.

3.6 Speech Synthesis and Generation

In human-computer interaction, the form of a computer response is as important as the content, and many applications include scenarios that require or are significantly enhanced by speech synthesis. For remote access to computers via phone 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 multi-media environment or simplify an application with many response windows by providing a non-visual 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 in the quality of current systems. In addition, advances in natural language generation open a new area of research, that is 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 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 [73]. 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.

Thus, the lack of funding in speech synthesis has impeded work in speech more generally, and has neglected the fact that communication via spoken language involves two participants, the speaker and the hearer. Neglecting one participant leads inevitably to compromised and frustrating communication, and continued neglect of synthesis will impede research on spoken language in human-machine interactions. Further, the recent work in response generation makes an even stronger case for the importance of work in synthesis.

3.6.1 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. Therefore, NSF funding of speech synthesis should include research on both text-to-speech and speech generation. 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 [62, 26].

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 [130, 132].

3. **Integration of synthesis and language generation.**

   Very 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 in speech generation. Possibilities range from simply increasing the quality to modeling mood 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.

3.6.2 Benefits of Research on Speech Synthesis

Speech output will be useful for many types of interactive systems, but the benefits are perhaps most clear in applications involving information access via phone, 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 [122, 90].

Speech synthesis also 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.

Finally, research in speech generation can enable more natural and more effective human-machine interaction. Use of synthesized responses can help guide the user with respect to system capabilities, as well as increase the communication bandwidth from a computer to the human user. In addition, speech generation allows telephone access to systems with spoken language understanding capabilities.

3.7 Multi-lingual Systems

At present, research in the United States in speech and natural language is almost exclusively aimed at monolingual communication in American English. Recent cataclysmic shifts in geo-politics 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 just as international telephone traffic is growing. The scientific community, though always somewhat international, is ever more so, as a result of which, 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 multi-lingual speech and language processing systems.

3.7.1 Key Research Challenges and Directions

The NSF research program, as distinct from programs sponsored by more mission-oriented agencies such as DARPA, should focus on fundamental scientific issues which will be important for a range of applications in multi-lingual 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. **Multi-lingual 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 data base retrieval or command and control of complex systems.

2. **Language Identification.**

   As an independent capability or as a part of a multi-lingual 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. **Multi-lingual Text and Speech Generation from Multi-modal Data Bases.**

   Complementary to the multi-lingual 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 [131, 17], 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.

3.7.2 Fundamental Scientific Issues

In the context of addressing the challenges delineated above, the NSF program should address fundamental scientific issues for automated multi-lingual spoken language systems. Examples of such issues include:

1. The general question of what are the fundamental acoustic, perceptual, and linguistic differences among languages that should be investigated, with a view toward accommodating these differences in multi-lingual systems.

2. 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?

3. 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.

4. 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.

5. 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?

6. 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 multi-lingual applications.

3.7.3 Benefits of Research on Multi-lingual Systems

The benefits of this research are described in Sections 5.3 and 5.5 below.

3.8 Interactive Multimodal Systems

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, for example, 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. 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 advanced interdisciplinary research.

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[80]." 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 judgements 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 [94, 95, 93, 137, 136]. Stork et al. trained separate neural networks to recognize spoken letters using (1) acoustic features only, (2) visual facial features only, or (3) combined acoustic and visual features. In comparison with the other two alternatives, the network based on combined visual and acoustic information performed more accurately and degraded more gracefully as ambient noise levels increased [119]. 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: (1) a choice of which information channel is used, or (2) converging sources of information from more than one channel.

In addition to concerns for increased robustness and utility, there is a great need for future systems that can support more natural, seamless, and flexible human-computer interaction. In this regard, multimodal systems that incorporate speech have the potential to demonstrate important inferface advantages. For example, well designed multimodal systems are expected to provide more flexibility with respect to input and output alternatives, to permit easier avoidance and correction of errors, to be more natural and easy to learn, and so forth.

3.8.1 Key Research Challenges and Fundamental Scientific Issues

In order to work toward the development of more facile and productive multimodal systems, ones capable of forging a complementary synthesis among different component modalities, 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 generate novel strategies for designing multimodal systems with performance characteristics superior to those of simpler unimodal alternatives. Among other things, the successful cultivation of such systems will require advance empirical work with human subjects, building a variety of new prototype systems, and the development of appropriate metrics for evaluating the accuracy, efficiency, learnability, expressive power, and other characteristics 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 either input or 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.**

   Due to the relative complexity of multimodal systems, developing appropriate methods for guiding their design is critically needed. 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.

3.8.2 Benefits of Research on 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 portables, will bring computing to a larger and more diverse user population than ever before.

4 Infrastructure

Spoken language processing is a field where it is particularly important to support the scientific infrastructure. The problems are inherently multi-disciplinary, 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. This requires an infrastructure for training researchers with the necessary skills, for establishing effective communication channels, and for providing the computer resources, algorithms, data, and tools needed to optimize productivity. Although cooperation will increase research productivity, it is important to encourage and support creativity and diversity across sites as well.

4.1 Multi-disciplinary Research and Training

We must ensure that young researchers are adequately supported and trained. First and foremost, NSF should continue to expand its graduate and undergraduate scholarships. It should also increase the support for post-doctoral fellowships, visiting scientists, and sabbatical leaves to various institutions, including those abroad.

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. NSF should specifically encourage research projects that involve multiple principal investigators, perhaps across institutional boundaries.

The multi-disciplinary 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. There are currently no programs that train researchers in the area of spoken language understanding. To fill this gap, we recommend that the NSF sponsor a Summer Spoken Language Institute, perhaps similar to the Summer Linguistics Institute. There is a clear demand for such a program: V. 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, and language understanding. 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 multi-disciplinary collaboration is particularly clear in the area of multimodal systems. 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. Since such teams must represent a span of disciplines, technologies, and often research sites, they may frequently require relatively large working groups, with two researchers as minimal, and three to six more typical. Cross-training of future researchers clearly becomes a particularly critical issue in the successful development of multimodal systems. Therefore, we would encourage NSF to consider supporting summer institutes for research professionals, industry-academic internships, and advanced training for graduate and postdoctoral-level fellows that focus on this purpose.

4.2 Post-doctoral funding

Due to the low cost of Ph.D. students and the relatively higher cost of post-doctoral candidates, in the past investigators have primarily requested funding for students. However, given the current economic climate, there are not as many Ph.D. students entering programs and students leaving are often expected to have had a post-doc before obtaining academic positions. Since there is also a need for intermixing of different disciplines as indicated in the above paragraph, another way to achieve this would be to provide postdoctoral fellowships that a graduating Ph.D. could apply for, that would provide research support not restricted in advance to a particular university or industry laboratory. This would allow postdocs to move to different universities or laboratories with their own funding in order to encourage multi-disciplinary projects. In fact, NSF could stipulate criteria for how the fellowships could be used that would encourage multi-disciplinary interaction.

4.3 Database Development and 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. 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 report. 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.

The collection, transcription and distribution of speech data for spoken language understanding is a massive task. Fortunately, DARPA has supported speech data collection and distribution of these data to the research community through NIST. Recently, as a pre-competitive initiative, the U.S. Congress has established the Linguistic Database Consortium, administered through DARPA, to expand this activity. NSF should strive to make the speech corpora that result from these activities available to scientists in a timely manner.

There are a number of important unresolved issues in database development, including transcription conventions, levels of description (sounds, words) and a reliable system for transcribing the prosodic structure of speech. NSF can support database development by supporting research in this area, by supporting workshops, by supporting data collection not performed in other programs, and by providing the necessary computer resources and tools.

Data collection and related infrastructure support must be extended to research in speech synthesis. NSF should fund data collection efforts in support of speech synthesis research, and include as an evaluation criterion whether/how the data be made available to other researchers.

4.4 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 NSF grant. Institutional support, particularly at universities, is often not sufficient to allow this level of replacement. Therefore more NSF support for workstation replacement would improve research productivity and be a very cost-effective investment. Some degree of hardware/software standardization would 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 are that whatever computational power is available 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 non-interactive 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. In fact, our choice of experimental algorithms is restricted because of computational efficiency so that the runs can be done in reasonable time. It would be preferable to broaden the search 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 NSF should be funding.

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 vs. 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. This is particularly true when the experiment involves real time interaction by voice with a machine. In such cases it is often necessary to use a special architecture designed expressly for the experimental purpose. Such machines are usually 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.

The most common compromise that is used for speech processing is to use a specialized accelerator attached to a general processor such as a workstation. The specialized subsystem typically has a large number of efficiently coded library routines, so that low level coding is less of a problem for common uses. Care must still be taken that the overhead to communicate with the accelerator does not swamp the performance gains.

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 KB/sec 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.

4.5 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.

As part of their 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 will be available to interested researchers in the near future [27]. If this program is successful, it should be expanded to provide software support for other areas of spoken language understanding.

The notion of a turnkey system for spoken language understanding research, consisting of both software tools and standard computer configuration, received enthusiastic support. We recommend that NSF support the development and deployment of such turnkey systems. A more general solution to this problem is further described in section 4.7 .

4.6 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 email 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. NSF can provide this infrastructure to accomplish this in a number of ways:

1. **Speech Email mailing list.**

   The "connectionist" mailing list (operated out of CMU) can serve as a case study of this medium. To avoid excessive costs, such mailing lists can be self maintaining, with only active researchers invited to subscribe and under acceptance of a code of "reasonable" behavior on the net. The role of such a network is to announce major new results, preprints of published or unpublished papers, or meetings, conferences and workshops. It can also be used to raise issues and discussion of major concern for everybody in the field.

Support of such a facility would be minimal, with only a fraction of a graduate student's support to maintain mailing lists and some modest computing resources required.

2. **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. 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). Once again the database is self-maintaining and requires minimal attention and resources: a fraction of a student's support and a fileserver with a suitably large disk.

3. **Facility for electronic benchmarking and other algorithms.**

   Such a facility would be a repository for common databases, 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.

4. **Video-conferencing, multi-media 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, 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 some day also enable distributed teams of researchers to work together on joint projects under joint funding.

4.7 US Science Institute (NSF Initiative)

The workshop discussed the formation of a US Science Institute to serve as a central clearinghouse for curriculum development and educational activities, as well as a communication focal point for scientists. In these capacities, it would also become a computing resource center, disseminating tools (hardware and software) and providing shared high-performance computing to sites across the country. Such an institute would aid research in many areas of science in the United States by making information available more quickly, minimize duplication of effort by research organizations, and provide appropriate resources on a short-term basis to individual research groups.

It is important to consider how such an institute would relate to existing institutions such as the Linguistic Data Consortium or the Consortium for Lexical Research, as well as institutions such as the National Institute for Standards and Technology and even NSF itself. The Science Institute is envisioned as a physical entity, with office space, computing resources, and facilities, in contrast with the consortia, which are primarily collaborative mechanisms, although each has a host site. Its focus would be to provide basic infrastructure to the scientific community, in particular for education (K-12 through post-doctoral education), but also for scientific interchange of many kinds: mailing lists, publications databases, software exchange, and support for long-distance collaborations. All of these needs have been discussed in previous subsections. The Science Institute provides a mechanism to co-ordinate these activities and to develop the infrastructure to support them.

5 Benefits

5.1 Student Education and Jobs

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 waiting for qualified leaders. There are few multi-disciplinary centers of excellence in spoken language understanding in the U. S., 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 multi-disciplinary 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,

5.2 U.S. Competitiveness in the Global Marketplace

"Technology propels our economy forward. Without doubt, it has been our strongest competitive advantage. Innovation has created whole new industries and the renewal of existing ones. State-of-the-art products have commanded premium prices in world markets and technological advances have spurred productivity gains. Thus, America owes much of its standard of living to U.S. preeminence in technology." (John Young, in a report to the president of the U.S. entitled "President's Commission on Industrial Competitiveness").

Young states that the United States is losing its ability to compete technologically in world markets. "We have failed to respond adequately. ... Even in high technology - often referred to as the 'sunrise' industries - the United States has lost world market share in 7 out of 10 sectors. ... In order to make technology a continuing competitive advantage for the United States, we need to do three basic things: (1) create a solid foundation of science and technology that is relevant to commercial uses; (2) apply advances in knowledge to commercial products and processes; and (3) protect intellectual property by strengthening patent, copyright, trademark and trade secret protections. Attaining these goals will require actions on the part of the Federal Government, industry and our Nation's universities."

Although this report was written in 1985, numerous reports have been written since that have confirmed the need for government attention to new technology development in order to maintain the industrial competitiveness of the United States. A look from within U.S. corporations shows heavy emphasis on finding ways to reduce costs, increase worker productivity and generate new revenue streams. As Snow et al. [115] states: " ... firms must adapt with increasing speed to market demands and competitors' innovations while simultaneously controlling and even lowering product or service costs..." The new equation links competitive success to doing fewer things, better, with less.

Spoken language systems are among the many technologies that will make a significant contribution to cost reduction and new revenue generation in American industry today. The computer industry (e.g. Apple Computer, IBM) , the financial industry (e.g., Citibank, American Express), and the telecommunications industry (e.g., AT&T, NYNEX, U.S. WEST) have made it clear in the formation of major R&D organizations devoted to speech systems development that this technology is important to their future competitive positions. Easy access to and control of information is key to each of these industries. Speech understanding systems allow ordinary telephones to act as database access terminals. Multi-billion dollar customer service operations can be front-ended with speech technology to reduce costs and offer new services cost effectively. Office and factory automation will make companies run more efficiently and more cost effectively. The American with Disabilities Act requires telecommunications companies to provide services to disadvantaged people; spoken language systems help to provide many of these services.

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 research advances for competitive advantage, provides the future leaders of the industry, and educates those who work in it.

5.3 International Cooperation and Business

Multi-lingual 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 open avenues for trade. Large corporations also stand to benefit. Many have installed and maintained laboratories and sales offices world-wide. Cross-language collaboration is becoming more and more commonplace. Overcoming communication barriers will decrease cost 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.

Within our own society, an understanding and technical means to overcoming language barriers may improve interaction among the many peoples within our own country. It may also raise an awareness of other cultures and languages, and we would hope to see a new generation of multi-lingually aware and adept students to emerge as a side effect. Finally, multi-lingual speech and language processing may aid in this educational process by way of easily accessible computer aided language instruction.

5.4 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 performing 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.

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 well 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 publics' access to information, from travel schedules to books. 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 in colloquial discourse or something similar to it. Making such information sources accessible to the general public would be a significant benefit of spoken language interfaces.

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. In a sense, enabling computers to recognize and understand speech would be a major boom to the computer industry because of the enormous variety of new applications in which these machines could be useful.

5.5 Benefit to Scientific Community

Spoken language systems will increase the productivity of researchers who work with computers on a daily basis. As multi-lingual 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 on the use of computers within the general population. Scientists and engineers in many disciplines will need to respond to this development.

6 Recommendations

1. NSF should provide an immediate increase in the level of funding for spoken language research, commensurate with the role of this technology in supporting national research priorities in information and communications technology. Only by increasing funding of basic research in human machine interaction will the United States be able to maintain its competitive advantage relative to Europe and Japan.

2. NSF funding should focus on basic research and new approaches to spoken language understanding. This report identifies a number of fundamental problem areas that currently receive little or no current funding. The resulting research will create advances in fundamental knowledge leading to technological breakthroughs. The funding of basic research and innovative approaches allows NSF to play an important complementary role to the DARPA speech project, which emphasizes advances in current technology rather than basic research.

3. NSF should encourage collaborative and interdisciplinary research in spoken language understanding. Solutions to the fundamental problems described in this report require expertise in signal processing, statistics and pattern matching, speech and hearing science, linguistics, psychology and computer science. NSF can encourage multi-disciplinary collaboration by funding multi-person grants, by supporting curriculum development for multi-disciplinary programs, including summer institutes, and by supporting the growth of centers of excellence in spoken language understanding through larger infrastructure grants.

4. NSF can provide infrastructure support to research in a number of ways, including (a) the development of an integrated set of research tools and databases, (b) a standard set of computer resources (workstation, storage, data capture and playback), (c) adoption of a funding strategy that encourages the distribution of research tools and hardware to laboratories at diverse sites, to make available research and training opportunities at many sites, (d) development of training materials for both undergraduate and graduate education in disciplines related to spoken language understanding, and (e) the establishment of a United States Science Institute, which would serve to co-ordinate ongoing activities in education, in dissemination of common resources, and in communication and collaboration among researchers.

5. NSF should fund a broad spectrum of approaches to spoken language understanding. It is premature to favor a particular technology or approach. Basic research, research on system development, and research on data collection and evaluation methodologies are all vital to our progress in this area.

Acknowledgement

The authors are grateful to Vince Weatherill of the Center for Spoken Language Understanding at the Oregon Graduate Institute for producing and mailing several drafts of the report, and for integrating the many contributions by different authors into a cohesive final document.

References

[1] J. Allen, S. Guez, L. Hoebel, E. Hinkelman, K. Jackson, A. Kyburg, and D. Traum. The discourse system project. Technical Report 317, Computer Science Department, University of Rochester, November 1989.

[2] J. F. Allen and C. R. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 3(15):143–178, 1980.

[3] D. Appelt and E. Jackson. SRI International February 1992 ATIS benchmark test results. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[4] D. C. Bateman, D. K. Bye, and M. J. Hunt. Spectral contrast normalization and other techniques for speech recognition in noise. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, pages I.241–244. IEEE, Mar 1992.

[5] M. Berger and H. F. Silverman. Microphone array optimization by stochastic region contraction (SRC). *IEEE Transactions on Signal Processing*, 39(11):2377–2386, 1991.

[6] S. F. Boll. Suppression of acoustic noise in speech using spectral substraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27:113–120, 1979.

[7] G. Brown, K. Currie, and J. Kenworthy. *Questions of Intonation.* Croom Helm, 1980.

[8] B. Butterworth. Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, 4:75–87, 1975.

[9] R. Carlson and B Granstrom. Speech synthesis development and phonetic research – a personal introduction. *Journal of Phonetics*, 19:3–8, 1991.

[10] R. Carlson and L. Nord. Positional variants of some Swedish sonorants in an analysis-synthesis scheme. *Journal of Phonetics*, 19:49–60, 1991.

[11] J. Cheshire. *English Around the World: Sociolinguistic Perspectives.* Cambridge University 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 Research*, 1:185–195, 1979.

[13] K. W. Church. Phonological parsing and lexical retrieval. *Cognition*, 25:54–69, 1987.

[14] J. R. Cohen. Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, 85:2623–2629, 1989.

[15] D. Van Compernolle, W. Ma, F. Xie, and M. Van Diest. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9(5/6):433–442, December 1990.

[16] Ed. D. E. Walker. *Understanding Spoken Language.* North Holland, New York, 1978.

[17] R. Sproat D. Roe, F. Pereira. Efficient grammar processing for a spoken language translation system. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, page I.213. IEEE, Mar 1992.

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

[19] S. Dobler, P. Meyer, and H. W. Ruehl. A robust connected-words recognizer. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, pages I.245–248. IEEE, Mar 1992.

[20] J. Edwards, M. Beckman, and J. Fletcher. The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89:369–382, 1991.

[21] L. D. Erman, F. Hayes-Roth, V. R. Lesser, and D. R. Reddy. The Hearsay II speech understanding system. *ACM Computing Surveys*, pages 213–253, 1980.

[22] F. Kubala et al. BBN BYBLOS and HARC February 1992 ATIS benchmark results. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[23] L. Hirschman et al. Multi-site data collection for a spoken language corpus. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[24] W. Ward et al. Speech recognition in open tasks. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[25] G. Fant. Stops in CV-syllables. In G. Fant, editor, *Speech Sounds and Features*. MIT Press, Cambridge, MA, 1973.

[26] G. Fant. What can basic research contribute to speech synthesis? *Journal of Phonetics*, 19:75–90, 1991.

[27] M. Fanty, J. Pochmara, and R. A. Cole. An interactive environment for speech recognition research. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October 12–16 1992.

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

[29] J. L. Flanagan. Use of acoustic filtering to control the beamwidth of steered microphone arrays. *Journal of the Acoustical Society of America*, 78(2):423–428, August 1985.

[30] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, and M. M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58–71, February 1991.

[31] J. E. Flege. Laryngeal timing and phonation onset in utterance-initial English stops. *Journal of Phonetics*, 10:177–192, 1982.

[32] C. A. Fowler. Production and perception of coarticulation among stressed and unstressed syllables. *Journal of Speech and Hearing Research*, 46:127–149, 1981.

[33] D. B. Fry. The dependence of stress judgements on vowel formant structure. In *Proceedings of the 5th International Congress of Phonetic Sciences*, pages 306–311, 1965.

[34] T. Gay. Mechanisms in the control of speech rate. *Phonetica*, 38:148–158, 1981.

[35] O. Ghitza. Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment. *Journal of Phonetics*, 16(1):109–124, 1988.

[36] C. Gobl. Voice source dynamics in connected speech. Technical Report 1.1988, Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, 1988.

[37] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520. I.E.E.E., 1992.

[38] F. Grosjean and J. P. Gee. Prosodic structure and spoken word recognition. *Cognition*, 25,:135–155, 1987.

[39] B. J. Grosz. Discourse analysis. In D.E. Walker, editor, *Understanding Spoken Language*, pages 235–268. North Holland, New York, 1978.

[40] B. J. Grosz and C. L. Sidner. Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204, 1986.

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

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

[43] S. Hamlet. Handedness and articulatory asymmetries in /s/ and /l/. *Journal of Phonetics*, 15:191–195, 1987.

[44] J. Hampshire and A. Waibel. The meta-pi network: Connectionist rapid adaptation for high-performance multi-speaker phoneme recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.* IEEE, April 1990.

[45] B. A. Hanson and T. H. Applebaum. Robust speaker-independent word recognition using static, dynamic, and acceleration features: Experiments with Lombard and noisy speech. In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, pages 857–860. IEEE, 1990.

[46] K. S. Harris. Vowel duration change and its underlying physiological mechanisms. *Language and Speech*, 21:354–361, 1978.

[47] H. Hermansky. Perceptual linear predictive PLP analysis for speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[48] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, pages I–121–I–124. IEEE, March 1992.

[49] D. Hindle. *The Social and Situational Conditioning of Phonetic Variation.* PhD thesis, Univerisity of Pennsylvania, 1980.

[50] H. G. Hirsch, P. Meyer, and H. W. Ruehl. Improved speech recognition using high-pass filtering of subband envelopes. In *Proceedings of 2nd European Conference on Speech Communication and Technology*, pages 413–416, Genova, Italy, Sep 1991.

[51] H. Hirschberg. Towards a redefinition of yes/no question. In *Association for Computational Linguistics*, 22nd annual meeting of the ACL, 1983. Stanford University, California.

[52] J. Hirschberg and B. Gross. Intonational features of local and global discourse. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[53] J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proceedings of the Fifth DARPA Workshop on Speech and Natural Language*, Feb 1992.

[54] 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. *Journal of Phonetics*, 20:241–251, 1992.

[55] H.-W. Hon and K.-F. Lee. Vocabulary learning and environment normalization in vocabulary-independent speech recognition. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, pages I 485–488. IEEE, March 1992.

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

[57] K. Johnson. The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88:642–654, 1990.

[58] B. H. Juang. Speech recognition in adverse environments. *Computer Speech and Language*, pages 275–294, 1991.

[59] S. J. Kaplan. Cooperative responses from a portable natural language query system. *Artificial Intelligence*, 19(2), 1982.

[60] I. Karlsson. Female voices in speech synthesis. *Journal of Phonetics*, 19:111–120, 1991.

[61] D. Klatt and L. Klatt. Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[62] D. H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 3(82):737–793, 1987.

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

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

[65] W. Labov. The social stratification of (r) in New York city department stores. In W. Labov, editor, *Sociolinguistic Patterns.* University of Pennsylvania Press, Philadelphia, 1972.

[66] W. Labov. Sources of inherent variation in speech. In J. S. Perkell and D. H. Klatt, editors, *Invariance and Variability in Speech Processes*, pages 402–423. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[67] I. Lehiste. An acoustic-phonetic study of internal open juncture. *Supplement to Phonetica*, 1959.

[68] W. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

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

[70] B. Lindblom. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35:1773–1781, 1963.

[71] C. Linde and J. Goguen. Structure of planning discourse. *Journal of Social Biol. Structure*, 1:219–251, 1978.

[72] P. Lockwood, J. Boudy, and M. Blanchet. Non-linear spectral subtraction (NSS) and Hidden Markov Models for robust speech recognition in car noise environments. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, pages I.265–268. IEEE, 1992.

[73] J. S. Logan, B. G. Greene, and D. B. Pisoni. Segmental intelligibility of synthetic speech produced by ten text-to-speech systems. *Journal of the Acoustical Society of America*, 86:566–581, 1986.

[74] R. F. Lyon and C. Mead. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1119–1134, 1988.

[75] D. D. MacDonald. Natural language generation as a computational problem: An introduction. In M. Brady and R.C. Berwick, editors, *Computational Models of Discourse.* M. I. T. Press, Cambridge, Mass., 1983.

[76] A. Malhotra. Design criteria for a knowledge-based English language system for management: an experimental analysis. Technical Report MAC TR-146, MIT, 1975.

[77] W. C. Mann and J. A. Moore. Computer generation of multiparagraph English text. *American Journal of Computational Linguistics*, 1(7), 1981.

[78] K. F. McCoy. The ROMPER system: Responding to object-related misconceptions using perspective. In *24th Annual Meeting of the ACL.* Association of Computational Linguistics, New York City, New York, June 1986.

[79] M. McCutcheon, A. Hasegawa, and S. Fletcher. Effects of palatal morphology on [s,z] articulation. *Journal of the Acoustical Society of America*, 67:S94, 1980.

[80] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[81] K. R. McKeown. *Text Generation.* Cambridge University Press, Cambridge, England, 1985.

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

[83] K.R. McKeown and W. R. Swartout. Language generation and explanation. In J.F. Traub et al., editor, *Annual Review of Computer Science.* Annual Reviews Inc., Palo Alto, Ca, 1987.

[84] D. Monsour and B. H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1659–1671, 1989.

[85] J. M. Mullenix, D. B. Pisoni, and C. S. Martin. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85:365–378, 1989.

[86] F. Nolan. *The Phonetic Basis of Speaker Recognition.* Cambridge University Press, 1983.

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

[88] S. E. G. Ohman. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39:151–168, 1966.

[89] D.J. Ostry and K. G. Munhall. Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, 77:640–648, 1985.

[90] Ed. P. Suppes. University-level computer-assisted instruction at Stanford: 1968-1980, 1981. Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, CA.

[91] D. Pallett. ATIS benchmarks. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[92] C. L. Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 3(14), 1988.

[93] A. Pentland and K. Mase. Lip reading: Automatic visual recognition of spoken words. In *Proceedings of Image Understanding and Machine Vision.* Optical Society of America, June 12-14 1989.

[94] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, pages 26–29, Atlanta, Georgia, November 1984.

[95] E. D. Petajan, B. Bischoff, and D. Bodoff. An improved automatic lipreading system to enhance speech recognition. In *Proceedings of the ACM SIGCHI-88*, pages 19–25, 1988.

[96] J. Pierrehumbert. Prosody, intonation, and speech technology. Cambridge University Press. In press.

[97] J. Pierrehumbert and D. Talkin. Lenition of /h/ and glottal stop. In G. Docherty and D. R. Ladd, editors, *Papers in Laboratory Phonology II: Gesture, Segment, Prosody.* Cambridge University Press, 1992.

[98] R. K. Potter, G. A. Kopp, and H.C. Green. *Visible Speech.* van Nostrand, New York, 1947.

[99] P. Price. Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, 8:261–277, 1989.

[100] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90:2956–2970, 1991.

[101] M. Randolph. *Syllable Based Constraints on Properties of English.* PhD thesis, MIT, 1989.

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

[103] B. Repp. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Pscyhological Bulletin*, 92:81–110, 1982.

[104] B. Repp. Coarticulation in sequences of two nonhomorganic stop consonants: Perceptual and acoustic evidence. *Journal of the Acoustical Society of America*, 74:420–427, 1983.

[105] R. Schwartz, Y. Chow, and F. Kubala. Rapid speaker adaption using a probabalistic spectral mapping. In *Proceedings of the 1987 International Conference on Acoustics, Speech and Signal Processing*, pages 633–636. IEEE, 1987.

[106] D. Scott and A. Cutler. Segmental phonology and the perception of syntactic structure. *Journal of Verbal Learning and Verbal Behavior*, 23:450–466, 1984.

[107] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76, 1988.

[108] S. Seneff. A relaxation method for understanding spontaneous utterances. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[109] Shortliffe. *Computer-Based Medical Consultations.* Elsevier, New York, 1976.

[110] E. Shriberg, J. Bear, and J. Dowding. Automatic detection and correction of repairs in computer-human dialog. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[111] H. F. Silverman. Some analysis of microphone arrays for speech data acquisition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(2):1699–1712, December 1987.

[112] H. F. Silverman and S. E. Kirtman. A two-stage algorithm for determining talker location from linear microphone-array data. *Computer, Speech, and Language*, 6(2):129–152, April 1992.

[113] H. F. Silverman, S. E. Kirtman, J. E. Adcock, and P. C. Meuse. Experimental results for baseline speech recognition performance using input acquired from a linear microphone array. In *Notebook of the Fifth DARPA Workshop on Speech and Natuaral Language*, Arden House, Harriman, NY, February 1992.

[114] R. W. Smith, D. R. Hipp, and A. W. Biermann. A dialogue control algorithm and its performance. *Third Conference on Applied Natural Language Processing*, March 31 - April 3 1992. Trento Italy.

[115] C.C. Snow, R.E. Miles, and H.J. Coleman Jr. Managing 21st Century network organizations. *Organizational Dynamics*, Winter, 1992.

[116] R. M. Stern, F. Liu, Y. Ohshima, T. M. Sullivan, and A. Acero. Multiple approaches to robust speech recognition. In *Notebook of the Fifth DARPA Workshop on Speech and Natuaral Language*, Arden House, Harriman, NY, February 1992.

[117] K. N. Stevens and A. S. House. Perturbation of vowel articulation by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 6:111–128, 1963.

[118] K.N. Stevens. *Evidence for the role of acoustic boundaries in the perception of speech sounds*, pages 243–255. Academic Press, New York, 1985.

[119] D. G. Stork, G. Wolff, and El Levine. Neural network lipreading system for improved speech recognition. In *Proceedings of the International Joint Conference on Neural Networks, Vol II*, pages 286–295, 1992.

[120] W. V. Summers. Effects of stress and final consonant voicing on vowel production: articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, 82:847–863, 1987.

[121] B. Sundheim. Overview of the third message understanding evaluation and conference. In *Proceedings of the Third Message Understanding Conference MUC-3*, San Mateo, CA, 1991. Morgan Kaufmann.

[122] P. Suppes. Current trends in computer assisted instruction. In M.C. Yovits, editor, *In Advances in Computers.* Academic Press, 1979.

[123] 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. *Journal of the Acoustical Society of America*, 90:1309–1325, 1991.

[124] W.R. Swartout. XPLAIN: a system for creating and explaining expert consulting systems. *Artificial Intelligence*, 3(2):285–325, 1983.

[125] T. M. Teary. Static, dynamic and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85:2088–2113, 1989.

[126] H. Tennant. Experience with the evaluation of natural language question answerers. Technical report, Univ. of Illinois, Urbana-Champaign, 1979.

[127] H. Traunmuller. Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, 69:1465–1475, 1981.

[128] D. van Bergem, L. Pols, and F. Koopmans van Beinum. Peceptual normalization of the vowels of a man and a child. *Speech Communication*, 7:1–20, 1988.

[129] E. Vatikiotis-Bateson and J. A. S. Kelso. Rhythm type and articulatory dynamics in English, French, and Japanese. *Journal of Phonetics*, 21, 1992. In press.

[130] N. Veilleux and M. Ostendorf. Probabilistic parse scoring based on prosodic phrasing. In *DARPA Workshop on Speech and Natural Language Processing*, February 1992.

[131] A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann, and J. Tebelskis. JANUS: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.* IEEE, May 1991.

[132] M. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries computer speech and language. unpublished, 1992.

[133] M. Witbrock and P. Haffner. Rapid connectionist speaker adaptation. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, pages I–453–456. IEEE, March 1992.

[134] U. Wolz, K. R. McKeown, and G. Kaiser. Automated tutoring in interactive environments: A task centered approach. *Journal of Machine Mediated Learning*, 1989.

[135] S. R. Young, A. G. Hauptmann, W. H. Ward, E. T. Smith, and P. Werner. High level knowledge sources in usable speech recognition systems. *Communications of the ACM*, 32(2):183–194, Feb 1989.

[136] B. P. Yuhas, Jr. M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, November 1989.

[137] B. P. Yuhas, Jr. M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1988.

[138] V. W. Zue. The use of speech knowledge in automatic speech recognition. In *Proceedings of the IEEE*, pages 1602–1615. IEEE, 1985.

[139] V.W. Zue and M. Laferriere. An acoustic study of medila /t,d/ in American English. *Journal of the Acoustical Society of America*, 66:1039–1050, 1979.