SRI International

The Role of Voice in Human-Machine Communication

Technical Note No. 538

January 1994

By: Philip R. Cohen, Senior Computer Scientist
Sharon L. Oviatt, Senior Research Psychologist
Computer Diaglogue Laboratory
Artificial Intelligence Center
Computing and Engineering Sciences Division

This work was supported in part by the National Science Foundation under Grant No. IRI-9213472.

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

The Role of Voice in Human-Machine Communication[*]

Philip R. Cohen
Sr. Computer Scientist
and
Sharon L. Oviatt
Sr. Cognitive Scientist
Computer Dialogue Laboratory
Articial Intelligence Center
SRI International

August 20, 1993

Summary

Optimism is growing that the near future will witness rapid growth in human-computer interaction using voice. System prototypes have recently been built that demonstrate speaker-independent real-time speech recognition and understanding of naturally spoken utterances moderately sized vocabularies (1000 to 2000 word), and larger-vocabulary speech recognition systems are on the horizon. Already, computer manufacturers are building speech recognition subsystems into their new product lines. However, before this technology will be broadly useful, a substantial knowledge base about human spoken language and performance during computer-based interaction needs to be gathered and applied. This chapter reviews application areas in which spoken interaction may play a significant role, assesses potential benefits of spoken interaction with machines, and attempts to compare voice with alternative and complementary modalities of human-computer interaction. The chapter also discusses information that will be needed to build a firm empirical foundation for future designing human-computer interfaces. Finally, it argues for a more systematic and scientific approach to understanding human language and performance with voice-interactive systems.

[*] The writing of this paper was supported in part by a grant from the National Science Foundation (number IRI-9213472) to SRI International. This paper will appear in *Human-Computer Interaction by Voice*, D. B. Roe and J. Wilpon (eds.), chapter 1, National Academy of Sciences Press, Washington, D. C., 1993.

1 Introduction

From the beginning of the computer era, futurists have dreamed of the conversational computer — a machine that we could engage in spoken natural language conversation. For instance, Turing's famous "test" of computational intelligence imagined a computer that could conduct such a fluent English conversation that people could not distinguish it from a human. However, despite prolonged research and many notable scientific and technological achievements, until recently there have been few human-computer dialogues, none of them spoken. This situation has begun to change, as steady progress in speech recognition and natural language processing technologies, supported by dramatic advances in computer hardware, has made possible laboratory prototype systems with which one can engage in simple question-answering dialogues. Although far from human-level conversation, this initial capability is generating considerable interest and optimism for the future of human-computer interaction using voice.

This paper aims to identify applications for which spoken interaction may be advantageous, to situate voice with respect to alternative and complementary modalities of human-computer interaction, and to discuss obstacles that exist to the successful deployment of spoken language systems because of the nature of spoken language interaction.

Two general sorts of speech input technology are considered. First, we survey a number of existing applications of speech *recognition* technologies, for which the system identifies the words spoken, but need not understand the meaning of what is being said. Second, we concentrate on applications that will require a more complete *understanding* of the speaker's intended meaning, examining future spoken dialogue systems. Finally, we discuss how such speech understanding will play a role in future human-computer interactions, particularly those involving the coordinated use of multiple communication modalities, such as graphics, handwriting, and gesturing. It is argued that progress has been impeded by the lack of adequate scientific knowledge about human spoken interactions, especially with computers. Such a knowledge base is essential to the development of well-founded human-interface guidelines that can assist system designers in producing successful applications incorporating spoken interaction. Given recent technological developments, the field is now in a position to systematically expand that knowledge base.

1.1 Background and Definitions

Human-computer interaction using voice may involve speech input or speech output, perhaps in combination with each other or with other modalities of communication.

1.1.1 Speech Analysis

The speech analysis task is often characterized along five dimensions:

Speaker Dependence. Speech recognizers are described as speaker-dependent/trained, speaker-adaptive, and speaker-independent. For speaker-dependent recognition, samples of a given user's speech are collected and used as models for his/her subsequent utterances. For speaker-adaptive recognition, parameterized acoustical models are initially available, which can be more finely tuned for a given user through pronunciation of a limited set of specified utterances. Finally, speaker-independent recognizers are designed to handle any user's speech, without training, in the given domain of discourse [45].

Speech Continuity. Utterances can be spoken in an isolated manner, with breaks between words, or as continous natural speech.

Speech Type. To develop initial algorithms, researchers typically first use read speech as data, in which speakers read random sentences drawn from some corpus, such as the Wall Street Journal. Subsequent to this stage of algorithm development, speech recognition research attempts to handle spontaneous speech, in which speakers construct new utterances in the chosen domain of discourse.

Interactivity. Certain speech recognition tasks, such as dictation, can be characterized as non-interactive, in that the speaker is receiving no feedback from the intended listener(s). Other systems are designed to process interactive speech, in which speakers construct utterances as part of an exchange of turns with a system or with another speaker.

Vocabulary and Grammar. The user can speak words from a tightly constrained vocabulary and grammar, or from larger vocabularies and grammars that more closely approximate that of a natural language. The system's vocabulary and grammar can be chosen by the system designer or application developer, or it can be compiled from data based on actual users speaking either to a simulated system or to an early system prototype. Current speech recognition technologies require an estimate of the probability of occurrence of each word in the context of the other words in the vocabulary. Because these probabilities are typically approximated from the distribution of words in a given corpus, it is currently difficult to expand a system's vocabulary, although research is proceeding on vocabulary-independent recognition [61].

Vendors often describe their speech recognition hardware as offering a very high recognition accuracy, but it is only in the context of a quantitative understanding of the recognition task that one can meaningfully compare the performance of recognizers. To calibrate the difficulty of a given recognition task for a given system, researchers have come to use a measure of the *perplexity* of that system's language model, which measures, roughly-speaking, the average number of word possibilities at each state of the grammar [8, 9, 67]. Word recognition accuracy has been found, in general, to be inversely proportional to perplexity. Most commercial systems offer speech recognition systems claiming >95% word recognition accuracy given a perplexity on the order of 10. At least one vendor offers a 1000-5000 word speaker-independent system, with perplexities in the range of 66-433, and a corresponding word-recognition error of 3% to 15% for recognition of isolated words [10]. Current laboratory systems support real-time speaker-independent recognition of continuously spoken utterances drawn from a vocabulary of approximately 1500 words, with a perplexity of 50-70, resulting in word recognition error rates between 4% and 8% [114]. The most ambitious speaker-independent systems are currently recognizing, in real-time, read speech drawn from a 5,000 word vocabulary of Wall Street Journal text, with a perplexity of 120, resulting in a word-recognition error rate of 5% [114]. Larger vocabularies are now being attempted.

The end result of voice recognition is the highest ranking string(s) of words, or often lattice of words, that covers the signal. For small vocabularies and tightly constrained grammars, a simple interpreter can respond to the spoken words directly. However, for larger vocabularies and more natural grammars, *natural language understanding* must be applied to the output of the recognizer in order to recover the intended meaning of the utterance.[1] Because this natural language

[1] See [97] for a discussion of how these components can be integrated.

understanding process is complex and open-ended, it is often constrained by the application task (e.g., retrieving information from a database) and by the domain of discourse (e.g., a database about airline flights). The combination of speech recognition and language understanding will be termed here *speech understanding*, and the systems that use such input will be termed *spoken language systems.* This chapter reviews earlier work on uses of speech recognition, but concentrates on uses of spoken language.

1.1.2 Speech Synthesis

Three forms of speech synthesis technology exist:

Digitized Speech. To produce an utterance, the machine assembles and plays back previously recorded and compressed samples of human speech. Although a noticeable break between samples can often be heard, and the overall intonation may be inaccurate, such a synthesis process can offer human-sounding speech of high intelligibility. This process is, however, limited to producing combinations of the recorded samples.

Text-to-Speech. Text-to-speech synthesis involves an automated analysis of the structure of words into their morphological constituents. By combining the pronunciations of those subword units according to letter- and morph-to-sound rules, coupled with a large list of exceptional pronunciations (for English), arbitrary text can be rendered as speech. Because this technology can handle open-ended text, it is suitable for large-scale applications such as reading text aloud to blind users or reading electronic mail over the telephone. Text-to-speech science and technology is covered at length in this volume [2, 20].

Concept-to-Speech. With text-to-speech systems, the text to be converted is supplied from a human source. Future dialogue systems will require computers to decide for themselves what to say and how to say it in order to arrive at a meaningful and contextually appropriate dialogue contribution. Such systems need to determine what speech action(s) to perform (e.g., request, suggestion), how to refer to entities in the utterance, what to say about them, what grammatical forms to use, and what intonation to apply. Moreover, the utterance should contribute to the course of the dialogue, so the system should keep a representation of what it has said in order to analyze and understand the user's subsequent utterances.

The research areas of speech synthesis and language generation have received considerably less attention than speech recognition and understanding, but will increase in importance as the possibility of developing spoken dialogue systems becomes realizable.

The remainder of this chapter explores current and future application areas in which spoken interaction may be a preferred modality of communication with computers. First, factors that may influence the desirability and efficiency of voice-based interaction with computers are identified, independent of whether a simple command language or a quasi-natural language is being spoken. Then, we discuss spoken language interaction, comparing it both to keyboard-based interaction and to the currently dominant graphical user-interface paradigm. After identifying circumstances that favor spoken language interaction, gaps in the scientific knowledge base of spoken communication are identified that present obstacles to the development of spoken language-based systems. It is observed that future systems will be multimodal, with voice being only one of the communication modalities available. We conclude with suggestions for further research that needs to be undertaken to support the development of voice-based unimodal and multimodal systems, and argue that there is a pressing need to create empirically-based human interface guidelines for system developers before voice-based technology can fulfill its potential.

2 When Is Spoken Interaction with Computers Useful?

As yet, there is no theory or categorization of tasks and environments that would predict, all else being equal, when voice would be a preferred modality of human-computer communication. Still, a number of situations have been identified in which spoken communication with machines may be advantageous:

- When the user's hands or eyes are busy
- When only a limited keyboard and/or screen is available
- When the user is disabled
- When pronunciation is the subject matter of computer use
- When natural language interaction is preferred

We briefly examine the present and future roles of spoken interaction with computers for these environments. Because spoken natural language interaction is the most difficult to implement, we discuss it extensively in Section 3.3.2.

2.1 Voice Input

2.1.1 Hand/Eyes-Busy Tasks

The classic situation favoring spoken interaction with machines is one in which the user's hands and/or eyes are busy performing some other task. In such circumstances, by using voice to communicate with the machine, people are free to pay attention to their task, rather than breaking away to use a keyboard. Field studies suggest that, for example, F-16 pilots who can attain a high speech recognition rate can perform missions, such as formation flying or low-level navigation, faster and more accurately when employing spoken control over various avionics subsystems, as compared with keyboard and multifunction-button data entry [62, 126, 155]. Similar results have been found for helicopter pilots in noisy environments during tracking and communications tasks [135, 136, 143].[2]

Commercial hands/eyes-busy applications also abound. For instance, wire installers, who spoke a wire's serial number and then were guided verbally by the computer to install that wire achieved a 20-30% speedup in productivity, with improved accuracy and lower training time, over their prior manual method of wire identification and installation [93]. Parcel sorters who spoke city names instead of typing destination-labeled keys attained a 37% improvement in entry time during hands/eyes busy operations [149]. However, when the hands/eyes-busy component of parcel sorting was removed, spoken input offered no distinct speed advantages. In addition, VLSI circuit designers were able to complete 24% more tasks when spoken commands were available than when they used

[2] Further discussion of speech recognition for military environments can be found in [152, 153].

only a keyboard and mouse interface (see Section 3.3.1) [94]. Although individual field studies are rarely conclusive, many field studies of highly accurate speech recognition systems with hands/eyes-busy tasks have found that spoken input leads to higher task productivity and accuracy.

Not only does spoken input offer efficiency gains for a given hands/eyes-busy task, it offers the potential to change the nature of that task in beneficial ways. For example, instead of having to remember and speak or type the letters "YYZ" to indicate a destination airport, a baggage-handler could simply say "Toronto," thereby employing an easy-to-remember name [94, 104]. Similar potential advantages are identified for voice-based telephone dialers, to which one can say "Call Tom," rather than having to remember and input a phone number [123]. Other hands/eyes busy applications that might benefit from voice interaction include data entry and machine control in factories and field applications [95], access to information for military command-and-control, astronauts' information management during extra-vehicular access in space, dictation of medical diagnoses [10], maintenance and repair of equipment, control of automobile equipment (e.g., radios, telephones, climate control), and navigational aids [142].

A major factor determining success for speech input applications is speech recognition accuracy. For example, the best task performance reported during AFTI F-16 test flights was obtained once pilots attained isolated word recognition rates > 95%. Below 90%, the effort needed to correct recognition errors was said to outweigh the benefits gained for the user [62]. Similar results showing the elimination of benefits once error correction is considered have also been found in tasks as simple as entry of connected digits [57].

To attain a sufficiently high level of recognition accuracy in field tests, spoken input has been severely constrained to allow only a small number of possible words at any given time. Still, even with such constraints, accuracy in the field often lags that of laboratory tests because of many complicating factors, such as the user's physical and emotional state, ambient noise, microphone equipment, the demands of real tasks, methods of user and system training, and individual differences encountered when an array of real users is sampled. However, it is claimed that most failures of speech technology have been the result of human factors engineering and management [81], rather than low recognition accuracy per se. Human factors issues are discussed further below, and in [69].

2.1.2 Limited Keyboard/Screen Option

The most prevalent current uses of speech synthesis and recognition are telephone-based applications. Speech synthesizers are commonly used in the telecommunications industry to support directory assistance, speaking the desired telephone number to the caller, thereby freeing the operator to handle another call. Speech recognizers have been deployed to replace or augment operator services (e.g., collect calls), handling hundreds of millions of callers each year, and resulting in multi-million dollar savings [83, 101, 156]. Speech recognizers for telecommunications applications accept a very limited vocabulary, perhaps spotting only certain key words in the input, but need to function with high reliability for a broad spectrum of the general public. Although not as physically severe as avionic or manufacturing applications, telecommunications applications are difficult because callers receive little or no training about use of the system, and may have low-quality equipment, noisy telephone lines, and unpredictable ambient noise levels. Moreover caller behavior is difficult to predict and channel [11, 69, 139]. [3]

The considerable success at automating the simpler operator services opens the possibility for more ambitious telephone-based applications, such as information access from remote databases. For example, the caller might inquire about airline and train schedules [1, 39, 116], yellow-pages information, or bank account balances [101], and receive the answer auditorily. This general area of human-computer interaction is much more difficult to implement than simple operator services because the range of caller behavior is quite broad, and because speech understanding and dialogue participation is required, rather than just word recognition. When even modest quantities of data need to be conveyed, a purely vocal interaction may be difficult to conduct, although the advent of "screen phones" may well improve such cases.

Perhaps the most challenging potential application of telephone-based spoken language technology is the interpretation of telephony [80, 125] in which two callers speaking different languages can engage in a dialogue mediated by a spoken language translation system [75, 160]. Such systems are currently designed to incorporate speech recognition, machine translation, and speech synthesis subsystems, and to interpret one sentence at a time. A recent initial experiment organized by ATR International (Japan), with Carnegie-Mellon University (USA) and Siemens A.G. (Germany), involved Japanese-English and Japanese-German machine-interpreted dialogues [120, 160]. Utterances in one language were recognized and translated by a local computer, which sent a translated textual rendition to the foreign site, where text-to-speech synthesis took place. AT&T has demonstrated a limited-domain spoken English-Spanish translation system [125], although not a telephone-based one, and Nippon Electric Corporation has demonstrated a similar Japanese-English system.

Apart from the use of telephones, a second equipment-related factor favoring voice-based interaction is the ever-decreasing size of portable computers. Portable computing and communications devices will soon be too small to allow for use of a keyboard, implying that the input modalities for such machines will most likely be digitizing pen and voice [36, 107], with screen and voice providing system output. Given that these devices are intended to supplant both computer and telephone, users will already be speaking *through* them. A natural evolution of the devices will offer the user the capability to speak *to* them as well.

Finally, an emerging use of voice technology is to replace the many control buttons on consumer electronic devices (e.g., VCRs, receivers). As the number of user-controllable functions on these devices increases, the user interface becomes overly complex and can lead to confusion over how to perform even simple tasks. Products have recently been announced that allow users to program their devices using simple voice commands.

2.1.3 Disability

A major potential use of voice technology will be to assist deaf users in communicating with the hearing world using a telephone [16]. Such a system would recognize the hearing person's speech, render it as text, and synthesize the deaf person's textual reply (if using a computer terminal) as a spoken utterance. Another use of speech recognition in assisting deaf users would be captioning television programs or movies in real time. Speech recognition could also be used by motorically impaired users to control suitably augmented household appliances, wheel chairs, and robotic pros-

[3] An excellent review of the human factors and technical difficulties encountered in telecommunications applications of speech recognition can be found in [70].

theses. Text-to-speech synthesis can assist users with speech and motor impediments, can assist blind users with computer interaction and, when coupled with optical character recognition technology, can read printed materials to blind users. Finally, given sufficiently capable speech recognition systems, spoken input may become a prescribed therapy for repetitive stress injuries, such as carpal tunnel syndrome, which are estimated to afflict approximately 1.5% of office workers in occupations that typically involve the use of keyboards [144], although speech recognizers may themselves lead to different repetitive stress injuries [92].[4]

2.1.4 Subject Matter is Pronunciation

Speech recognition will become a component of future computer-based aids for foreign language learning and for the teaching of reading [17, 18, 98]. For such systems, speakers' pronunciation of computer-supplied texts would be analyzed and given as input to a program for teaching reading or foreign languages. Whereas these may be easier applications of speech recognition than some because the words being spoken are supplied by the computer, the recognition system will still be confronted with mispronunciations and slowed pronunciations, requiring a degree of robustness not often considered in other applications of speech recognition. Substantial research will also be needed to develop and field-test new educational software that can take advantage of speech recognition and synthesis for teaching reading. This is perhaps one of the most important potential applications of speech technology because the societal implications of raising literacy levels on a broad scale are enormous.

2.2 Voice Output

As with speech input, the factors favoring voice output are only informally understood. Just as tasks with a high degree of visual or manual activity may be more effectively accomplished using spoken input, such tasks may also favor spoken system output. A user could concentrate on a task rather than altering his or her gaze to view a system display. Typical application environments include flying a plane, in which the pilot could receive information about the status of the plane's subsystems during critical phases of operation (e.g., landing, high-speed maneuvering), and driving a car, in which the driver would be receiving navigational information in the course of driving. Other factors thought to favor voice output include remote access to information services over the telephone, lack of reading skills, darkened environments, and the need for omnidirectional information presentation, as in the issuing of warnings in cockpits, control rooms, factories, etc. [136, 146].

There are numerous studies of speech synthesis, but no clear picture has emerged of when computer-human communication using speech output is most effective or preferred. Psychological research has investigated the intelligibility, naturalness, comprehensibility, and recallability of synthesized speech (e.g., [89, 103, 136, 146]). Intelligibility and naturalness are orthogonal dimensions, in that synthetic speech present in an environment of other human voices may be intelligible, but unnatural. Conversely, human speech in a noisy environment may be natural, but unintelligible [136]. Many factors influence the intelligibility of synthesized speech in an actual application environment, including the baseline phoneme intelligibility, speaking rate, signal-to-noise level, and presence of other competing voices, as well as the linguistic and pragmatic context [136, 137].

[4] The general subject of "assistive technology" is covered at length elsewhere in this volume [86], and a survey of speech recognition for rehabilitation can be found in [16].

The desirability of voice output depends on the application environment. Pilots prefer to hear warnings with synthetic speech rather than digitized speech, as the former is more easily distinguished from other voices, such as radio traffic [150]. However, in simulations of air traffic control systems, in which pilots would expect to interact with a human, digitized human speech was preferred to computer synthesized speech [136]. Users may prefer to receive information visually, either on a separate screen or on a heads-up display [143], reserving spoken output for critical warning messages [136]. Much more research is required in order to determine those types of information processing environments for which spoken output is beneficial and preferred. Furthermore, rather than just concentrating on the benefits of speaking an utterance as compared with other modes of presenting the same information, future research needs to evaluate user performance and preferences as a function of the *content* of what is being communicated, especially if the computer will be determining that content (e.g., the generation of navigational instructions for drivers). Finally, research is critically necessary to develop algorithms for determining the appropriate intonation contours to use during a spoken human-computer dialogue.

2.3 Summary

There are numerous existing applications of voice-based human-computer interaction, and new opportunities are developing rapidly. In many applications for which the user's input can be constrained sufficiently to allow for high recognition accuracy, voice input has been found to lead to faster task performance with fewer errors than keyboard entry. Unfortunately, no principled method yet exists to predict when voice input will be the most effective, efficient, or preferred modality of communication. Similarly, no comprehensive analysis has identified the circumstances when voice will be the preferred or most efficient form of computer output, though again hands/eyes busy tasks may also be among the leading candidates for voice output.

One important circumstance favoring human-computer communication by voice is when the user wishes to interact with the machine in a natural language, such as English. The next section discusses such spoken language communication.

3 Comparison of Spoken Language with Other Communication Modalities

A user who will be speaking to a machine may expect to be able to speak in a natural language, that is, to use ordinary linguistic constructs such as noun and verb phrases. Conversely, if natural language interaction is chosen as a modality of human-computer communication, users may prefer to speak rather than type. In either case, users may expect to be able to engage in a dialogue, in which each party's utterance sets the context for interpreting subsequent utterances. We first discuss the status of the development of spoken language systems, and then compare spoken language interaction with typed interaction.

3.1 Spoken Language System Prototypes

Research is progressing on the development of spoken language question-answering systems — systems that allow users to speak their questions freely, and which then understand those questions and provide an accurate reply. The ARPA-supported air-travel information systems [1], developed at

Bolt, Beranek and Newman [79], Carnegie-Mellon University [63], Massachusetts Institute of Technology [165], SRI International [7], and other institutions, allow novice users to obtain information in real-time from the Official Airline Guide database, through the use of speaker-independent, continuously spoken English questions. The systems recognize the words in the user's utterance, analyze the meaning of those utterances, often in spite of word recognition errors, retrieve information from (a subset of) the Official Airline Guide's database, and produce a tabular set of answers that satisfy the question. These systems respond with the correct table of flights for over 70% of context-independent questions, such as "Which flights depart from San Francisco for Washington after 7:45 a.m.?" Rapid progress has been made in the development of these systems, with a 4-fold reduction in weighted error-rates recognition over a 20-month period for speech recognition, a 3.5-fold reduction over a 30-month period for natural language understanding, and a 2-fold reduction over a 20-month period for their combination as a spoken language understanding system. Other major efforts to develop spoken dialogue systems are also ongoing in Europe [91, 116], and Japan [160].

Much of the language processing technology used for spoken language understanding has been based on techniques for keyboard-based natural language systems.[5] However, spoken input presents qualitatively different problems for language understanding that have no analogue in keyboard interaction.

3.2 Spoken Language vs. Typed Language

3.2.1 Research Methodology

In our review of findings about linguistic communication relevant to spoken human-computer interaction, some results are based on analyses of human-human interaction, some are based on human-to-simulated-computer interaction, and some are based on human-computer interaction. Studies of human-human communication can identify the communicative capabilities that people bring to their interactions with computers, and can show what could be achieved, were computers adequate conversationalists. However, because this level of conversational competence will be unachievable for some time, scientists have developed techniques for simulating computer systems that interact via spoken language [5, 46, 51, 54, 82, 112, 113, 115, 122] by using a concealed human assistant who responds to the spoken language. With this method, researchers can analyze people's language, dialogue, task performance, and preferences, before developing fully functional systems.

Important methodological issues for such simulations include providing accurate and rapid response, and training the simulation assistant to function appropriately. Humans engage in rapid spoken interaction, and bring expectations for speed to their interaction with computers. Slow interactions can cause users to interrupt the system with repetitions while the system is processing their earlier input [148] and, it is conjectured, also can elicit phenomena characteristic of noninteractive speech [111]. One technique used to speed up such voice-in/voice-out simulations is the use of a vocoder, which transforms the assistant's naturally spoken response into a mechanical-sounding utterance [46, 54]. The speed of the "system" is thus governed by the assistant's knowledge and reaction time, as well as the task at hand, but not by speech recognition, language understanding, and speech synthesis. However, because people speak differently to a computer than they do to a person [46], even to prompts for simple yes/no answers [11, 12], the assistant should not provide *too*

[5] For a discussion of the state of research and technology of natural language processing, see [13, this volume].

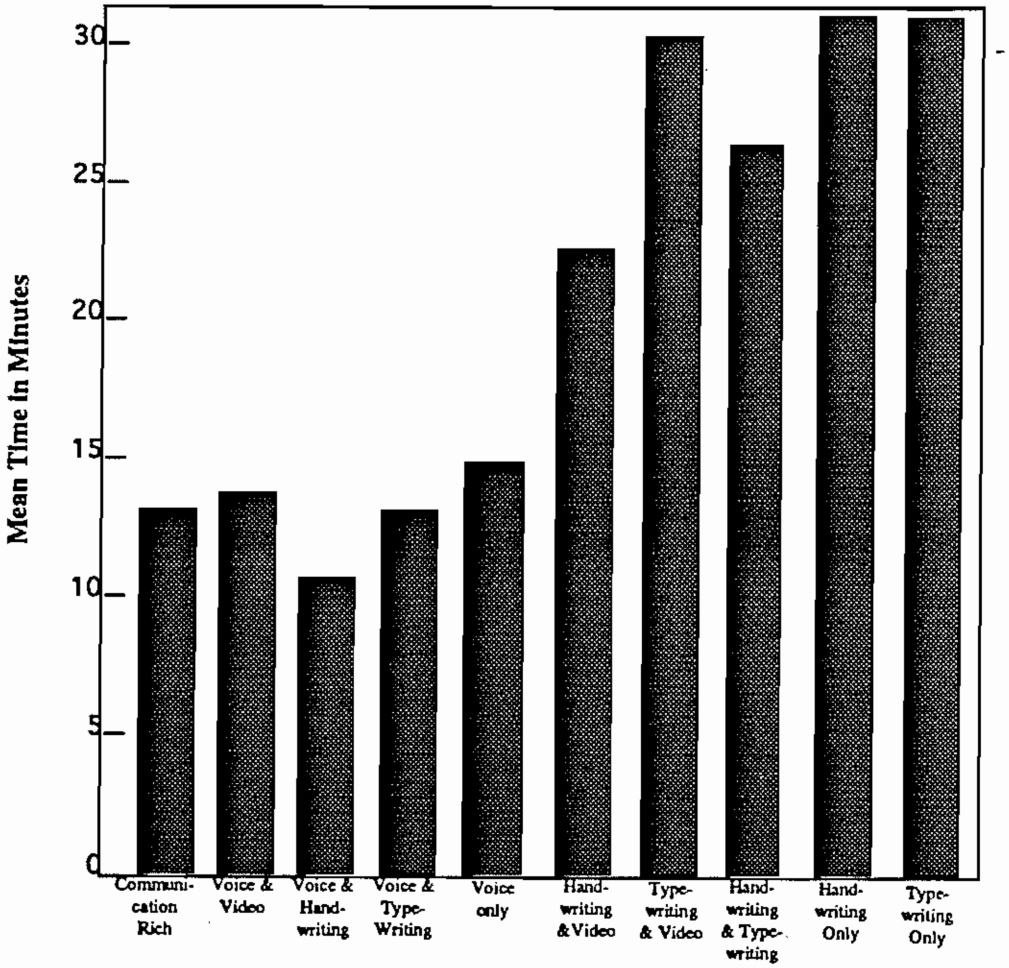
intelligent a reply, as this might reveal the "system" as a simulation. A second simulation method, which both constrains the simulation assistant and supports a rapid response, is to provide the assistant with certain pre-defined fields and structures on the screen that can be selected to reply to the subject [5, 38, 82, 112]. More research is needed into the development of simulation methodologies that can accurately model spoken language systems, such that patterns of interaction with the simulator are predictive of interaction patterns with the actual spoken language system.

3.2.2 Comparison of Language-based Communication Modalities

In a series of studies of interactive human-human communication, Chapanis and colleagues ([22, 23, 73, 96, 105]) compared the efficiency of human-human communication when subjects used any of 10 communication modalities (including face-to-face, voice-only, linked teletypes, interactive handwriting). The most important determinant of a team's problem-solving speed was found to be the presence of a voice component. Specifically, a variety of tasks were solved 2- to 3-times faster using a voice modality than a hardcopy one, as illustrated in Figure 1. At the same time, speech led to an 8-fold increase in the number of messages and sentences, and a 10-fold increase in rate of communicating words. These results indicate the substantial *potential* for efficiency advantages that may result from use of spoken language communication.

Research by the authors confirmed these efficiency results in human-human dialogues to perform equipment assembly tasks [28, 110], finding a 3-fold speed advantage for interactive telephone speech over keyboard communication. Furthermore, the structure of telephone dialogues differed from that of keyboard dialogues. Among the difference, spoken dialogues exhibited more cue phrases that signaled the structure of the dialogue (such as "next", "ok now"), and speakers interacted in a more "fine-grained" fashion than did keyboard users. Specifically, in order to achieve a subtask, speakers often made two requests, one for object identification and one for action, whereas keyboard users typically integrated both into one imperative utterance. Similar findings of a fine-grained approach during spoken interaction versus a more syntactically integrated approach for keyboard interaction have been found in a study of simulated human-computer interaction [164]. Finally, spoken input was more "indirect" than keyboard input. That is, unlike keyboard interaction, spoken utterances did not literally convey the speaker's intention that the listener perform an action [28]. Future research needs to address the extent to which such results generalize to spoken human-computer interaction for comparable tasks.

One benefit of voice input is the elimination of typing, which could offer potential office productivity savings [10, 68]. In a study of a simulated "listening typewriter," Gould et al. [49, 50, 51] examined how novice and expert users of dictation would use a machine that could recognize and type the user's dictation of a business letter, as compared with dictating and editing the letter to a human, or handwriting and editing the letter. The listening typewriter system was simulated, and the subjects were informed that they were in fact speaking to a person. It was claimed that users of a listening typewriter were as satisfied with that mode of communication as with the others, and that dictating to a listening typewriter could potentially be as fast a mode of letter composition as typing. There is, however, countervailing evidence from a number of simulation studies [99, 102] that speech-only word processors are less efficient and less preferred than composition methods based on writing or typing. Moreover, a combined method of using speech for text input and touch screen for cursor control was more efficient than speech alone, though still less efficient than composition and editing using keyboards or handwriting.



Mean problem solution times for the 10 modes of communication

Figure 1: Voice determines task efficiency (from Ochsman and Chapanis [105]).

Neither series of studies examined in detail the linguistic and discourse structure of the dictated material that might explain why spoken composition and editing is less efficient than other modalities. In a study of human-human communication, it was found that inexperienced "dictators" providing instructions for a human listener produced more discourse structures that would require editing in order to make acceptable text, such as repetitions, elaborations, and unusual uses of referring expressions, than did users of interactive speech or interactive keyboard [110, 111]. Thus, lack of interaction with a listener may contribute to poorly formulated input, placing a larger burden on the post-editing phase where speech input is less efficient [102]. In summary, though automatic dictation devices have been much touted as an important product concept for speech technology, their potential benefit remains a question.

The space of modality studies has not yet been systematically explored. We do not know precisely how results from human-human communication studies can predict results for studies of human-simulation or human-computer interactions. Also, more studies comparing the structure and content of spoken human-computer language with typed human-computer language need to be conducted in order to understand how to adapt technology developed for keyboard interaction to spoken language systems.

Common to many successful applications of voice-based technology is the lack of an adequate alternative to voice, given the task and environment of computer use. Major questions remain as to the applications where voice will be favored when other modalities of communication are possible. Some studies report a decided preference for speech when compared to other modalities [128], yet other studies report an opposite conclusion [99, 102]. Thus, despite the aforementioned potential benefits of human-computer interaction using voice, it is not obvious why people should want to speak to their computers in performing their daily office work. To provide a framework for answering this question, the discussion below compares the currently dominant direct-manipulation user interface with typed or spoken natural language.

3.3 Comparison of Natural Language Interaction with Alternative Modalities

Numerous alternative modalities of human-computer interaction exist, such as the use of keyboards for transmitting text, pointing and gesturing with devices such as the mouse, a digitizing pen, trackballs, touchscreens, and digitizing gloves. It is important to understand what role speech, and specifically spoken language, can play in supporting human interaction, especially when these other modalities are available. To begin this discussion, we need to identify properties of successful interfaces. Such an interface should ideally be:

Error free. The interface should prevent the the user from formulating erroneous commands, should minimize misinterpretations of the user's intent, and should offer simple methods for error-correction.

Transparent. The functionality of the application system should obvious to the user.

High-level. The user should not have to learn the underlying computer structures and languages, but rather should be able to state simply his or her desires, and have the system handle the details.

Consistent. Strategies that work for invoking one computer function should transfer to the invocation of others.

Easy to Learn. The user should not need formal training, but rather a brief process of exploration should suffice for learning how to use a given system.

Expressive. The user should be able to perform easily any combination of tasks in mind, within the bounds of the system's intended functionality.

Using this set of properties, we discuss the use of direct manipulation and natural language technologies.

3.3.1 Direct Manipulation

The graphical user-interface paradigm involves a style interaction that offers the user menus, icons, and pointing devices (e.g., the "mouse" [42]) to invoke computer commands, as well as multiple windows in which to display the output. These graphical user interfaces (GUIs), popularized by the Apple Macintosh and by Microsoft Windows, employ techniques pioneered at SRI International and at Xerox's Palo Alto Research Center in the late 1960s and 1970s [41, 72]. With GUIs, users perform actions by selecting objects and then choosing the desired action from a menu, rather than by typing commands.

In addition, with many GUIs a user can directly manipulate graphical objects in order to perform actions on the objects the represent. For example, a user can copy a file from one disk to another by selecting its icon with the pointing device, and "dragging" it from the list of files on the first disk to the second. Other direct manipulation actions include using a "scroll bar" to view different sections of a file, and dragging a file's icon on top of the "trash" icon to delete it. Apart from the mouse, numerous pointing devices exist, such as trackballs and joysticks, and some devices offer multiple capabilities, such as the use of pens for pointing, gesturing, and handwriting. Finally, to generalize along a different dimension, users now can directly manipulate virtual worlds using computer-instrumented gloves and body-suits [44, 78, 124], allowing for subtle effects of body motion to affect the virtual environment.

Strengths. Many writers have identified virtues of well-designed graphically-based direct manipulation interfaces (DMIs) (e.g., [64, 132]), claiming that

- Direct manipulation interfaces based on familiar metaphors are intuitive and easy to use.
- Graphical user interfaces can have a consistent "look and feel" that enables users of one program to learn another program quickly.
- Menus make the available options clear, thereby curtailing user errors in formulating commands and specifying their arguments.
- GUIs can shield the user from having to learn underlying computer concepts and details.

It is no exaggeration to say that graphical user interfaces supporting direct manipulation interaction have been so successful that no serious computer company would attempt to sell a machine without one.

Weaknesses. Direct manipulation interfaces do not suffice for all needs. One clear expressive weakness is the paucity of means available for identifying entities. Merely allowing users to select currently displayed entities provides them little support for identifying objects not on the screen (such as a file name in a list of 200 files), for specifying temporal relations that denote future or past events, for identifying and operating on large sets of entities, and for using the context of interaction. At most, developers of GUIs have provided simple string-matching routines that find objects based on exact or partial matches of their names. What is missing is a way for users to *describe* entities using some form of linguistic expression in order to denote or pick out an individual object, a set of objects, time period, and so forth.[6] At a minimum, a description language should include some way to find entities having a given set of properties, to say which properties are of interest as well as which properties are not, to say how many entities are desired, to supply temporal constraints on actions involving those properties, and so forth. Moreover, a useful feature

[6] Of course, the elimination of descriptions was a conscious design decision by the originators of GUIs.

of a description language is the ability to reuse the referents of previous descriptions. Some of these capabilities are found in formal query languages, and all are found in natural languages.

Although shielding a user from implementation details, direct manipulation interfaces are often not high-level. For example, one common way to request information from a relational database is to select certain fields from tables that one wants to see. To do this correctly, the user needs to learn the structure of the database — for example, that the data is represented in one or more tables, comprised of numerous fields, whose meanings may not be obvious. Thus, the underlying tabular implementation has become the user interface metaphor. An alternative is to develop systems and interfaces that translate between the user's way of thinking about the problem and the implementation. In so doing, the user might perhaps implicitly retrieve information, but need not know that it is kept in a database, much less learn the structure of that database. By engaging in such a high-level interaction, users may be able to combine information access with other information processing applications, such as running a simulation, without first having to think about database retrieval, and then switching "applications" mentally to think about simulation.

When numerous commands are possible, GUIs usually present a hierarchical menu structure. As the number of commands grows, the casual user may have difficulty remembering in which menu they are located. However, the user who knows where the desired action is located in a large action hierarchy still needs to navigate the hierarchy. Software designers have attempted to overcome this problem by providing different menu sets for users of different levels of expertise, by preselecting the most recently used item in a menu, and by providing direct links to commonly used commands through special key combinations. However, in doing the latter, GUIs are borrowing from keyboard-based interfaces and command languages.

Because direct manipulation emphasizes rapid, graphical response to actions [132], the time of system action in DMIs is literally the time at which the action was invoked. Although some systems can delay actions until specific future times, DMIs and GUIs offer little support for users who want to execute actions at an unknown but describable future time.

Finally, DMIs rely heavily on a user's hands and eyes. Given our earlier discussion, certain tasks would be better performed with speech. So far, however, there is little research comparing graphical user interfaces with speech. Early laboratory results of a direct-manipulation VLSI design system augmented with speaker-dependent speech recognition indicate that users were as fast at speaking single-word commands as they were at invoking the same comands with mouse-button clicks, or by typing a single letter command abbreviation [94]. That is, no loss of efficiency occurred due to use of speech for simple tasks at which DMIs typically excel. Note that a 2- to 3-fold advantage in speed is generally found when speaking is compared to typing full words [23, 110]. In a recent study of human-computer interaction to retrieve information from a small database (of 240 entries), it was found that speech was substantially preferred over direct-manipulation use of scrolling, even though the overall time to complete the task with voice was longer [128]. This study suggests that, for simple risk-free tasks, user preference may be based on time-to-input rather than overall task completion times or overall task accuracy.

3.3.2 Natural Language Interaction

Strengths. Natural language is the paradigmatic case of an expressive mode of communication. A major strength is the use of psychologically salient and mnemonic descriptions. English, or any other natural language, provides a set of finely-honed descriptive tools such as the use of noun phrases for identifying objects, verb phrases for identifying events, and verb tense and aspect for describing time periods. By the very nature of sentences, these capabilities are deployed simultaneously, as sentences must be about something, and most often describe events situated in time.

Coupled with this ability to describe entities, natural languages offer the ability to avoid extensive redescription through the use of pronouns and other "anaphoric" expressions. Such expressions are usually intended to denote the same entities as earlier ones, and the recipient is intended to infer the connection. Thus, the use of anaphora provides an economical benefit to the speaker, at the expense of the listener's having to draw inferences.

Furthermore, natural language commands can offer a direct route to invoking an action or making selections that would be deeply embedded in the hierarchical menu of actions or would require multiple menu selections, such as font, type style and size in a word-processing program. In using such commands, a user could avoid having to select numerous menu entries to isolate the desired action. Moreover, because the invocation of an action may involve a description of its arguments, information retrieval is intimately woven into the invocation of actions.

Ideally, natural language systems should require only a minimum of training on the domain covered by the target system. Using natural language, people should be able to interact immediately with a system of known content and functionality, without having to learn its underlying computer structures. The system should have sufficient vocabulary, as well as linguistic, semantic, and dialogue capabilities, to support interactive problem solving by casual users — that is, users who infrequently employ the system. For example, at its present state of development, many users can successfully solve trip-planning problems with one of the ATIS systems [1], within a few minutes of introduction to the system and its coverage. To develop systems with this level of robustness, the system be trained and tested on a substantial amount of data representing input from a broad spectrum of users.[7] Currently, the level of training required to achieve a given level of proficiency in using these systems is unknown.

Weaknesses. In general, various disadvantages are apparent when natural language is incorporated into an interface. Pure natural language systems suffer from opaque linguistic and conceptual coverage — the user knows the system cannot interpret every utterance, but does not know precisely what it *can* interpret [58, 99, 138, 147]. Often, multiple attempts must be made to pose a query or command that the system can interpret correctly. Thus, such systems can be error-prone and, some claim [130], lead to frustration and disillusionment. One way to overcome these problems was suggested in a menu-based language processing system in which users composed queries in a quasi-natural language by selecting phrases from a menu [145]. Although the resulting queries are guaranteed to be analyzable, when there is a large number of menu choices to make, the query process becomes cumbersome.

Many natural language sentences are ambiguous, and parsers often find more ambiguities than people do. Hence, a natural language system often engages in some form of clarification or confirmation subdialogue to determine if its interpretation is the intended one. Current research is attempting to handle the ambiguity of natural language input by developing probabilistic parsing algorithms for which analyses would be ranked by their probability of occurrence in the given domain [90]. Also, research is beginning to investigate the potential for using prosody to choose

[7] The ATIS effort has required the collection and annotation of over 10,000 user utterances, some of which is used for system development, and the rest for testing during comparative evaluations conducted by the National Institute of Standards and Technology.

among ambiguous parses [15, 121]. A third research direction involves minimizing ambiguities through multimodal interface techniques to channel the user's language [29, 31, 113].

Another disadvantage of natural language interaction is that reference resolution algorithms do not always supply the correct answer, in part because systems have underdeveloped knowledge bases, and in part because the system has little access to the discourse situation, even if the system's prior utterances and graphical presentations have created that discourse situation. To complicate matters, systems currently have difficulty following the context shifts inherent in dialogue. These contextual and world knowledge limitations undermine the search for referents, and provide another reason that natural language systems are usually designed to confirm their interpretations.

It is not clear where typed natural language interaction will be a modality of choice. Studies comparing typed natural language database question answering with database querying using an artificial query language (e.g., SQL) [21] have given equivocal results, with some studies concluding that natural language interaction offers faster and more compact query formulation [66], while others conclude that database querying using SQL is more accurate and easier to learn [66, 131]. However, these studies are flawed by the use of prototype natural language systems rather than product-quality systems. When a product-quality natural language database retrieval system (INTELLECT [55]) was studied in the field, users reported efficiency gains and a clear preference for natural language interaction as compared with a previous query language method of database interaction [19]. Another difficulty in many laboratory studies is the lack of adequate controls on subject training. In one study comparing the utility of natural versus query language usage for database access [131], users in the natural language condition were given virtually no training on the content of a database, with the rationale that natural language systems should require no training, while users of SQL were trained on the file and field names of that database. Not surprisingly, under these conditions natural language users made more "overshoot" errors, in the sense of asking for information not contained in the database.

3.4 Summary: Circumstances Favoring Spoken Language Interaction with Machines

Theoretically, direct manipulation should be beneficial when the objects to be manipulated are on the screen, their identity is known, and there are not too many objects from which to select. In addition, graphical user interfaces limit users' options, preventing them from making errors in formulating commands. Natural language interaction with computers offers potential benefits when users need to identify objects, actions, and events from sets too large to be displayed and/or examined individually, and when users need to invoke actions at future times that must be described. Furthermore, natural language allows users to think about their problems and express their goals in their own terms rather than those of the computer. However, in allowing users to do so, systems need to have sufficient reasoning and interpretive capabilities to solve the problems of translating between the user's conceptual model and the system's implementation.

Combining the empirical results on circumstances favoring voice-based interaction with the foregoing analysis of interactions for which natural language may be most appropriate, it appears that applications requiring speedy user input of complex descriptions will favor spoken natural language communication. Moreover, this preference is likely to be stronger when a minimum of training about the underlying computer structures is possible. Examples of such an application area are asking questions of a database, or creating rules for action (e.g., "If I am late for a meeting, notify the meeting participants"). Because of the recency of usable spoken language systems, there are very few studies comparing spoken language interaction with direct manipulation for accomplishing real tasks.

So far, we have contrasted spoken interaction with other modalities. It is worth noting that these modalities have complementary advantages and disadvantages, which can be leveraged to develop multimodal interfaces that compensate for the weaknesses of one interface technology via the strengths of another [29, 31]. We discuss multimodal systems in section 5.

4 Human-Factors Obstacles to Spoken Language Systems

Although there are numerous technical challenges to building spoken language systems, many of which are detailed in this volume, interface and human-factors knowledge is especially needed about such systems. We consider below information needed about spontaneous speech, spoken natural language, and spoken interaction.

4.1 Spontaneous Speech

When an utterance is *spontaneously spoken*, it may well involves false starts, hesitations, filled pauses, repairs, fragments, and other types of technically "ungrammatical" utterances. These phenomena disrupt both speech recognizers and natural language parsers, and must be detected and corrected before techniques based on present technology can be deployed robustly. Current research has begun to investigate techniques for detecting and handling disfluencies in spoken human-computer interaction [14, 59, 100], and robust processing techniques have been developed that enable language analysis routines to recover the meaning of an utterance in spite of recognition errors [40, 63, 65, 140].

Assessment of different types of human-human and human-computer spoken language has revealed that people's rate of spontaneous disfluencies and self-repairs is substantially lower when they speak to a system, rather than another person [108]. A strong predictive relation has also been demonstrated between the rate of spoken disfluencies and an utterance's length [108]. Rather than having to resolve disfluencies, interface research has revealed that form-based techniques can reduce up to 70% of all disfluencies that occur during human-computer interaction [108]. In short, research suggests that some difficult types of input, such as disfluencies, may be avoided altogether through strategic interface design.

4.2 Natural Language

In general, because the human-machine communication in spoken language involves the system's understanding a *natural language*, but not the entire language, users will employ constructs outside the system's coverage. However, it is hoped that given sufficient data on which to base the development of grammars and templates, the likelihood will be small that a cooperative user will generate utterances outside the coverage of the system. Still, it is not currently known:

- How to select relatively "closed" domains, whose vocabulary and linguistic constructs can be acquired through iterative training and testing on a large corpus of user input

- How well users can discern the system's communicative capabilities

- How well users can stay within the bounds of those capabilities
- What level of task performance users can attain
- What level of misinterpretation users will tolerate, and what level is needed for them to solve problems effectively
- How much training is acceptable

Systems are not adept at handling linguistic coverage problems, other than responding that given words are not in the vocabulary, or that the utterance was not understood. Even recognizing that an out-of-vocabulary word has occurred is itself a difficult issue [35]. If users can discern the system's vocabulary, we can be optimistic that they can adapt to that vocabulary. In fact, human-human communication research has shown that users communicating by typing can solve problems as effectively with a constrained task-specific vocabulary (500 to 1000 words) as with an unlimited vocabulary [73, 96]. User adaption to vocabulary restrictions has also been found for simulated human-computer interaction [162, 164], although these results need to be verified for spoken human-computer interaction.

For interactive applications, the user may begin to *imitate* or *model* the language observed from the system, and the opportunity is present for the system to play an active role in *shaping* or *channeling* the user's language to match that coverage more closely. Numerous studies of human communication have shown that people will adopt the speech styles of their interlocutors, including vocal intensity [154], dialect [48], and tempo [141]. Explanations for this convergence of dialogue styles include social factors such as the desire for approval [48], and psycholinguistic factors associated with memory limitations [84]. Similar results have been found in a study of typed and spoken communication to a simulated natural language system [162, 163], which showed that people will model the vocabulary and length of the system's responses. For example, if the system's responses are terse, the user's input is more likely to be terse as well. In a simulation study of typed natural language database interactions, subjects modeled simple syntactic structures and lexical items that they observed in the system's paraphrases of their input [82]. However, it is not known if the modeling of syntactic structures occurs in *spoken* human-computer interaction. If users of spoken language systems do learn to adopt the grammatical structures they observe, then new forms of user training may be possible by having system designers adhere to the principle that any messages supplied to a user must be analyzable by the system's parser. One way to guarantee such system behavior would be to require the system to generate its utterances, rather than merely reciting canned text, employing a bi-directional grammar. Any utterances the system could generate using that grammar would thus be guaranteed to be parseable.

A number of studies have investigated methods for shaping user's language into the system's coverage. For telecommunications applications, the phrasing of system prompts for information spoken over the telephone dramatically influences the rate of caller compliance for the expected words and phrases [11, 127, 139]. For systems with screen-based feedback, human spoken language can be effectively channeled through the use of a form that the user fills out with speech [113]. Form-based interactions reduce the syntactic ambiguity of the user's speech by 65%, measured as the number of parses per utterance, thereby leading to user language that is simpler to process. At the same time, for the service transactions analyzed in this study, users were found to prefer forms-based spoken and written interaction over unconstrained ones by a factor of 2-to-1. Thus, not only can people's language be channeled, there appear to be cases where they prefer the guidance and sense of completion provided by a form.

4.3 Interaction and Dialogue

When given the opportunity to interact with systems via spoken natural language, users will attempt to engage in dialogues, expecting prior utterances and responses to set a context for subsequent utterances, and expecting their conversational partner to make use of that context to determine the referents of pronouns. Although pronouns and other context-dependent constructs sometimes occur less frequently in dialogues with machines than they do in human-human dialogues [74], context-dependence is nevertheless a cornerstone of human-computer interaction. For example, contextually-dependent utterances comprise 44% of the ATIS corpus collected for the ARPA spoken language community [53]. In general, a solution to the problem of understanding context-dependent utterances will be difficult, as it may require the system to deploy an arbitrary amount of world knowledge [24, 157]. However, it has been estimated that a simple strategy for referent determination employed in text processing, and one that uses only the syntactic structure of previous utterances, can suffice to identify the correct referent for pronouns in over 90% of cases [60]. Whether such techniques will work as well for spoken human-computer dialogue is unknown. One way to mitigate the inherent difficulty of referent determination when using a multimodal system may be to couple spoken pronouns and definite noun phrases with pointing actions [29, 31].

Present spoken language systems have supported dialogues in which the user asks multiple questions, some of which request further refinement of the answers to prior questions [1, 39], or dialogues in which the user is prompted for information [4, 116]. Much more varied dialogue behavior is likely to be required by users, such as the ability to engage in advisory, clarificatory and confirmatory dialogues [26, 87]. With respect to dialogue confirmations, spoken communication is tightly interactive and speakers expect rapid confirmation of understanding through backchannels (e.g, "uh huh") and other signals. Studies have shown that communication delays as brief as 0.25 second can disrupt conversation patterns [76], leading speakers to elaborate and rephrase their utterances [77, 111], and that telephone communications are especially sensitive to delays. The need for timely confirmations will challenge most applications of spoken language processing, particularly those involving telephony.

To support a broader range of dialogue behavior, more general models of dialogue are being investigated, both mathematically and computationally, including plan-based models of dialogue and dialogue grammars. Plan-based models are founded on the observation that utterances are not simply strings of words, but rather are the observable performance of communicative actions, or speech acts [129], such as requesting, informing, warning, suggesting, and confirming. Moreover, humans do not just perform actions randomly, but rather they plan their actions to achieve various goals, and in the case of communicative actions, those goals include changes to the mental states of listeners. For example, speakers' requests are planned to alter the intentions of their addressees. Plan-based theories of communicative action and dialogue [3, 6, 32, 34, 117, 134] assume that the speaker's speech acts are part of a plan, and the listener's job is to uncover and respond appropriately to the underlying plan, rather than just to the utterance. For example, in response to a customer's question of "Where are the steaks you advertised?", a butcher's reply of "How many do you want?" is appropriate because the butcher has discovered that the customer's plan of getting steaks himself is going to fail. Being cooperative, he attempts to execute a plan to achieve the customer's higher-level goal of having steaks [27]. Current research on this model is attempting to incorporate more complex dialogue phenomena, such as clarifications [88, 159, 87], and to model dialogue more as a *joint* enterprise, something the participants are doing together [25, 33, 52].

The dialogue grammar approach models dialogue simply as a finite state transition network [37, 119, 158], in which state transitions occur on the basis of the type of communicative action that has taken place (e.g., a request). Such automata might be used to predict the next dialogue "states" that are likely, and thus could help speech recognizers by altering the probabilities of various lexical, syntactic, semantic, and pragmatic information [4, 161]. However, a number of drawbacks to the model are evident [85, 30]. First, it requires that the communicative action(s) being performed by the speaker in issuing an utterance be identified, which itself is a difficult problem, for which prior solutions have required plan recognition [3, 71, 117]. Second, the model assumes that only one state results from a transition. However, utterances are multifunctional. An utterance can be, for example, both a rejection and an assertion. The dialogue grammar subsystem would thus need to be in multiple states simultaneously, a property typically not allowed. Finally, and most importantly, the model does not say how systems should choose amongst the next moves, i.e., the states currently reachable, in order for it to play its role as a cooperative conversant. Some analogue of planning is thus also likely to be required.

Dialogue research is currently the weakest link in the research program for developing spoken language systems. First and foremost, dialogue technology is in need of a specification methodology, in which a theorist could state formally what a dialogue system should *do* (i.e., what would count as acceptable dialogue behavior). As in other branches of computer science, such specifications may then lead to methods for mathematically and empirically evaluating whether a given system has met the specifications. However, to do this will require new theoretical approaches. Second, more implementation experiments need to be carried out, ranging from the simpler state-based dialogue models to the more comprehensive plan-based approaches. Research aimed at developing computationally tractable plan-recognition algorithms is critically needed.

5 Multimodal Systems

There is little doubt that voice will figure prominently in the array of potential interface technologies available to developers. Except for conventional telephone-based applications, however, human-computer interfaces incorporating voice will probably be multimodal, in the sense of combining voice with screen feedback use of a pointing device, gesturing, handwriting, etc. [31, 56, 107, 151]. Many application systems require multimodal communication, such as inherently map-based interactions. Such systems can involve coordinated speaking, gesturing, pointing, or writing on the map during input, and speech synthesis coordinated with graphics for output. From the previous discussion, it is apparent that each interface technology has strengths and weaknesses, and it may be strategic to attempt to develop interfaces that capitalize on the strengths of one to overcome weaknesses in another [29]. That is, users should be able to speak when desired, supplemented with other modalities as needed.

There are many advantages to multimodal interfaces:

Error Avoidance and Robust Performance. Multimodal interfaces can offer the potential to avoid errors that otherwise would be made in a unimodal interface. For example, it is estimated that 86% of the task-critical human performance errors that occurred during in a study of a interpreted telephony could have been avoided by opening up a screen-based handwriting channel [109]. Multimodal recognition also offers the possibility of enhanced recognition in adverse conditions. For example, simultaneous use of lip-reading speech recognizers may increase the recognition rate in high noise environments [47, 118] that otherwise would impair acoustic speech recognizers. Alternatively, in such environments, users of multimodal interfaces could simply switch modes, for example, to use handwriting.

Error Correction. Multimodal interfaces offer more options for correcting errors that do occur. Recognition errors present a problem to users, partly because their source is not apparent. Users frequently respond to speech recognition errors by hyperarticulating. But since recognizers are typically not trained on hyperarticulated speech, this repair strategy leads to a lower likelihood of successful recognition for that content [133]. Recognition problems can thus repeat numerous times on the same content, leading to a "degradation spiral" that is frustrating to users and may cause them to abort the application [107]. By providing the option of using another modality, such as handwriting, a user can simply switch modes in order to correct an error in the first modality.

Situational and User Variation. The various circumstances in which portable computers will be used is likely to alter people's preferences for one modality of communication or another. For example, the user may at times encounter noisy environments or desire privacy, and would therefore rather not speak. Also, people may prefer to speak for some task content, but prefer not to speak for others. Finally, different types of users may systematically prefer to use one modality rather than another. In all these cases, a multimodal system offers the needed flexibility.

Even as we investigate multimodal interaction for potential solutions to problems arising in speech-only applications, many implementation obstacles need to be overcome in order to integrate and synchronize modalities. For example, multimodal systems could present information graphically or in multiple coordinated modalities [43, 151], and permit users to refer linguistically to entities introduced graphically [29, 151]. Techniques need to be developed to synchronize input from simultaneous data streams, so that, for example, gestural inputs can help resolve ambiguities in speech processing, and vice versa. Research on multimodal interfaces needs to examine not only the techniques for forging a productive synthesis among modealities, but also the effect that specific integration architectures will have on human-computer interaction. Much more empirical research on the human use of multimodal systems needs to be undertaken, as yet we know relatively little about how people use multiple modalities in communicating with other people, let alone with computers, or about how to support such communication most effectively.

6 Scientific Research on Communication Modalities

The present research and development climate for speech-based technology is more active than it was at the time of the 1984 National Research Council report on speech recognition in severe environments [106]. Significant amounts of research and development funding are now being devoted to building speech understanding systems, and the first speaker-independent, continuous, real-time spoken language systems have been developed. However, some of the same problems identified then still exist today. In particular, few answers are available on how people will interact with systems using voice, and how well they will perform tasks in the target environments as opposed to the laboratory. There is little research on the dependence of communication on the modality used, or types of tasks, in part because there have not been principled taxonomies or comprehensive research addressing these factors. In particular, the use of multiple communication modalities to support human-computer interaction is only now being addressed.

Fortunately, the field is now in a position to fill gaps in its knowledge base about spoken human-machine communication. Using existing systems that understand real-time, continuously spoken utterances, which allow users to solve real problems, a number of vital studies can now be undertaken, in a more systematic manner. Examples include:

- Longitudinal studies of users' linguistic and problem-solving behavior that would explore how users adapt to a given system
- Studies of users' understanding of system limitations, and of their performance in observing the system's bounds
- Studies of different techniques for revealing a system's coverage, and for channeling user input
- Studies comparing the effectiveness of spoken language technology with alternatives, such as the use of keyboard-based natural language systems, query languages, or existing direct manipulation interfaces
- Studies analyzing users' language, task performance, and preferences to use different modalities, individually and within an integrated multimodal interface

The information gained from such studies would be an invaluable addition to the knowledge base of how spoken language processing can be woven into a usable human-computer interface. Sustained efforts need to be undertaken to develop more adequate spoken language simulation methods, to understand how to build limited but robust dialogue sytems based on a variety of communication modalities, and to study the nature of dialogue.

A vital and underappreciated contribution to the successful deployment of voice technology for human-computer interaction will come from the development of a principled and empirically-validated set of human-interface guidelines for interfaces that incorporate speech (cf. [81]). Graphical user-interface guidelines typically provide heuristics and suggestions for building "usable" interfaces, though often without basing such suggestions on scientifically established facts and principles. Despite the evident success of such guidelines for graphical user interfaces, it is not at all clear that a simple set of heuristics will work for spoken language technology, because human language is both more variable and creative than the behavior allowed by graphical user interfaces. Answers to some of the questions posed earlier would be valuable in laying a firm empirical foundation for developing effective guidelines for a new generation of language-oriented interfaces.

Ultimately, such a set of guidelines embodying the results of scientific theory and experimentation should be able to predict, given a specified communicative situation, task, user population, and a set of component modalities, what the user-computer interaction will be like with a multimodal interface of a certain configuration. Such predictions could inform the developers in advance about potential trouble spots, and could lead to a more robust, usable, and satisfying human-computer interface. Given the complexities of the design task, and the considerable expense required to create spoken language applications, if designers are left to their intuitions, applications will suffer.

Thus, for scientific, technological, and economic reasons, a concerted effort needs to be undertaken to develop a more scientific understanding of communication modalities, and how they can best be integrated in support of successful human-computer interaction.

7 Acknowledgements

Many thanks to Jared Bernstein, Clay Coler, Carol Simpson, Ray Perrault, Robert Markinson, Raja Rajasekharan, and John Vester for valuable discussions and source materials.

References

[1] Advanced Research Projects Agency. *ARPA Spoken Language Systems Technology Workshop*, Cambridge, Massachusetts, 1993. Massachusetts Institute of Technology.

[2] J. Allen. Linguistic aspects of speech synthesis. In D. B. Roe and J. Wilpon, editors, *Human-machine communication by voice*. National Academy of Sciences Press, Washington, D. C., 1993.

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

[4] F. Andry. Static and dynamic predictions: A method to improve speech understanding in cooperative dialogues. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October 1992. University of Alberta.

[5] F. Andry, E. Bilange, F. Charpentier, K. Choukri, M. Ponamalé, and S. Soudoplatoff. Computerised simulation tools for the design of an oral dialogue system. In *Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218)*. Commission of the European Communities, 1990.

[6] D. Appelt. *Planning English Sentences*. Cambridge University Press, Cambridge, U. K., 1985.

[7] D. E. Appelt and E. Jackson. SRI International February 1992 ATIS benchmark test results. In *Fifth DARPA Workshop on Speech and Natural Language*, San Mateo, Calif., 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.

[8] L. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.

[9] J. F. Baker. Stochastic modeling for automatic speech understanding. In D. R. Reddy, editor, *Speech Recognition*, pages 521–541. Academic Press, New York, 1975.

[10] J. M. Baker. Large vocabulary speaker-adaptive continuous speech recognition research overview at Dragon systems. In *Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology*, pages 29–32, Genova, Italy, 1991.

[11] S. Basson. Prompting the user in ASR applications. In *Proceedings of COST232 Workshop — European Cooperation in Science and Technology*, November 1992.

[12] S. Basson, O. Christie, S. Levas, and J. Spitz. Evaluating speech recognition potential in automating directory assistance call completion. In *AVIOS Proceedings.* American Voice I/O Society, 1989.

[13] M. Bates. Models of natural language understanding. In D. B. Roe and J. Wilpon, editors, *Human-machine communication by voice.* National Academy of Sciences Press, Washington, D. C., 1993.

[14] J. Bear, J. Dowding, and E. Shriberg. Detection and correction of repairs in human-computer dialog. In D. Walker, editor, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, June 1992.

[15] J. Bear and P. Price. Prosody, syntax and parsing. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 17–22, Pittsburgh, Pennsylvania, 1990.

[16] J. Bernstein. Applications of speech recognition technology in rehabilitation. In J. E. Harkins and B. M. Virvan, editors, *Speech to text: Today and tomorrow*, Washington, D. C., 1988. Gallaudet University Research Institute. GRI Monograph Series B., No. 2.

[17] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. Automatic evaluation and training in English pronunciation. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, pages 1185–1188, Kobe, Japan, 1990. The Acoustical Society of Japan.

[18] J. Bernstein and D. Rtischev. A voice interactive language instruction system. In *Proceedings of Eurospeech '91*, pages 981–984, Genova, Italy, 1991. IEEE.

[19] R. A. Capindale and R. G. Crawford. Using a natural language interface with casual users. *International Journal of Man-Machine Studies*, 32:341–362, 1990.

[20] R. Carlson. Models of speech synthesis. In D. B. Roe and J. Wilpon, editors, *Human-machine communication by voice.* National Academy of Sciences Press, Washington, D. C., 1993.

[21] D. D. Chamberlin and R. F. Boyce. Sequel: A structured english query language. In *Proceedings of the 1974 ACM SIGMOD Workshop on Data Description, Access and Control*, May 1974.

[22] A. Chapanis, R. B. Ochsman, R. N. Parrish, and G. D. Weeks. Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem solving. *Human Factors*, 14:487–509, 1972.

[23] A. Chapanis, R. N. Parrish, R. B. Ochsman, and G. D. Weeks. Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving. *Human Factors*, 19(2):101–125, April 1977.

[24] E. Charniak. Jack and Janet in search of a theory of knowledge. In *Advance Papers of the Third Meeting of the International Joint Conference on Artificial Intelligence*, Los Altos, California, August 1973. William Kaufmann Inc.

[25] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986. Reprinted in: *Intentions in Communication*, P. R. Cohen and J. Morgan and M. E. Pollack (eds.), MIT Press, Cambridge, Massachusetts, 1990.

[26] E. F. Codd. Seven steps to rendezvous with the casual user. In *Proceedings IFIP TC-2 Working Conference on Data Base Management Systems*, pages 179–200, Amsterdam, 1974. North-Holland Publishing Co.

[27] P. R. Cohen. *On Knowing what to Say: Planning Speech Acts.* PhD thesis, University of Toronto, Toronto, Canada, January 1978. Technical Report No. 118, Department of Computer Science.

[28] P. R. Cohen. The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10(2):97–146, April-June 1984.

[29] P. R. Cohen. The role of natural language in a multimodal interface. In *The 2nd FRIEND21 International Symposium on Next Generation Human Interface Technologies*, Tokyo, Japan, November 1991. Institute for Personalized Information Environment. Also appears in Proceedings of UIST'92, ACM Press, New York, 1992, 143-149.

[30] P. R. Cohen. Models of dialogue. In M. Nagao, editor, *Cognitive Processing for Vision and Voice: Proceedings of the Fourth NEC Research Symposium.* SIAM, 1993.

[31] P. R. Cohen, M. Dalrymple, D. B. Moran, F. C. N. Pereira, J. W. Sullivan, R. A. Gargan, J. L. Schlossberg, and S. W. Tyler. Synergistic use of direct manipulation and natural language. In *Human Factors in Computing Systems: CHI'89 Conference Proceedings*, pages 227–234, New York, New York, April 1989. ACM, Addison Wesley Publishing Co.

[32] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[33] P. R. Cohen and H. J. Levesque. Confirmations and joint action. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pages 951–957, Sydney, Australia, August 1991. Morgan Kaufmann Publishers, Inc.

[34] P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979.

[35] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Bierman, M. Bush, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorain, and V. Zue. NSF workshop on spoken language understanding. Technical Report CS/E 92-014, Oregon Graduate Institute, September 1992.

[36] H. D. Crane. Writing and talking to computers. Business Intelligence Program Report D91-1557, SRI International, Menlo Park, California, July 1991.

[37] N. Dahlbäck and A. Jönsson. An empirically based computationally tractable dialogue model. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society (COGSCI-92)*, Bloomington, Indiana, July 1992.

[38] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz studies — why and how. In L. Ahrenberg, N. Dahlbäck, and A. Jönsson, editors, *Proceedings from the Workshop on Empirical Models and Methodology for Natural Language Dialogue Systems*, Trento, Italy, April 1992. Association for Computational Linguistics, Third Conference on Applied Natural Language Processing.

[39] *Proceedings of the Speech and Natural Language Workshop*, San Mateo, California, October 1991. Morgan Kaufmann, Publishers, Inc.

[40] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran. Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 54–61, Columbus, Ohio, June 1993.

[41] D. Englebart. Design considerations for knowledge workshop terminals. In *National Computer Conference*, pages 221–227, 1973.

[42] W. K. English, D. C. Englebart, and M. A. Berman. Display-selection techniques for text manipulation. *IEEE Transactions on Human Factors in Electonics*, HFE-8(1):5–15, March 1967.

[43] S. K. Feiner and K. R. McKeown. COMET: Generating coordinated multimedia explanations. In *Human Factors in Computing Systems (CHI'91)*, pages 449–450, New York, April 1991. ACM-SIGCHI, ACM Press.

[44] S. Fisher. Virtual environments, personal simulation, and telepresence. *Multimedia Review: The Journal of Multimedia Computing*, 1(2), 1990.

[45] J. L. Flanagan. Overview of speech communication. In D. B. Roe and J. Wilpon, editors, *Human-machine communication by voice*. National Academy of Sciences Press, Washington, D. C., 1993.

[46] N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99, 1991.

[47] O. N. Garcia, A. J. Goldschen, and E. D. Petajan. Feature extraction for optical speech recognition or automatic lipreading. Technical report, Institute for Information Science and Technology, Department of Electrical Engineering and Computer Science, The George Washington University, Washington, D.C., November 1992.

[48] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson. Speech accommodation theory: The first decade and beyond. In M. L. McLaughlin, editor, *Communication Yearbook 10*, pages 13–48. Sage Publishers, Beverly Hills, California, 1987.

[49] J. D. Gould. How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4):648–661, 1978.

[50] J. D. Gould. Writing and speaking letters and messages. *International Journal of Man-Machine Studies*, 16(1):147–171, 1982.

[51] J. D. Gould, J. Conti, and T. Hovanyecz. Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26(4):295–308, April 1983.

[52] B. Grosz and C. Sidner. Plans for discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press, Cambridge, Massachusetts, 1990.

[53] MADCOW Working Group. Multi-site data collection for a spoken language corpus. In *Proceedings of the Speech and Natural Language Workshop*, pages 7–14, San Mateo, California, February 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.

[54] M. Guyomard and J. Siroux. Experimentation in the specification of an oral dialogue. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems.* Springer Verlag, Berlin, B. R. D., 1988. NATO ASI Series, vol. 46.

[55] R. Harris. User oriented data base query with the robot natural language query system. *International Journal of Man-Machine Studies*, 9:697–713, 1977.

[56] A. G. Hauptmann and P. McAvinney. Gestures with speech for direct manipulation. *International Journal of Man-Machine Studies*, 38:231–249, 1993.

[57] A. G. Hauptmann and A. I. Rudnicky. A comparison of speech and typed input. In *Proceedings of the Speech and Natural Language Workshop*, pages 219–224, San Mateo, California, June 1990. Morgan Kaufmann, Publishers, Inc.

[58] G. G. Hendrix and B. A. Walter. The intelligent assistant. *Byte*, pages 251–258, December 1987.

[59] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, Cambridge, Mass., June 1983.

[60] J. R. Hobbs. Resolving pronoun reference. *Lingua*, 44, 1978. Reprinted in *Readings in Natural Language Processing*, Grosz, B. J., Sparck Jones, K., and Webber, B. L. eds., Morgan Kaufman Publishers, Inc., Los Altos, California, 1986.

[61] H.-W. Hon and K.-F. Lee. Recent progress in robust vocabulary-independent speech recognition. In *Proceedings of the Speech and Natural Language Workshop*, pages 258–263, San Mateo, California, October 1991. Morgan Kaufmann, Publishers, Inc.

[62] J. A. Howard. Flight testing of the AFTI/F-16 voice interactive avionics system. In *Proceedings of Military Speech Tech 1987*, pages 76–82, Arlington, Virginia, 1987. Media Dimensions.

[63] X. Huang, F. Alleva, M.-Y. Hwang, and R. Rosenfeld. An overview of the SPHINX-II speech recognition system. In *Proc. of the ARPA Workshop on Human Language Technology*, San Mateo, Calif., 1993. Advanced Research Projects Agency, Morgan Kaufmann, Publishers, Inc.

[64] E. L. Hutchins, J. D. Hollan, and D. A. Norman. Direct manipulation interfaces. In D. A. Norman and S. W. Draper, editors, *User Centered System Design*, pages 87–124. Lawrence Erlbaum Publisher, Hillsdale, New Jersey, 1986.

[65] E. Jackson, D. Appelt, J. Bear, R. Moore, and A. Podlozny. A template matcher for robust NL interpretation. In *Proc. of the 4th DARPA Workshop on Speech and Natural Language*, pages 190–194, San Mateo, CA, February 1991. Morgan Kaufmann.

[66] M. Jarke, J. A. Turner, E. A. Stohr, Y. Vassiliou, N. H. WHite, and K. Michielsen. A field evaluation of natural language for data retrieval. *IEEE Transctions on Software Engineering*, SE-11(1):97–113, 1985.

[67] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532–536, April 1976.

[68] F. Jelinek. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, 73(11):1616–1624, November 1985.

[69] C. A. Kamm. User interfaces for voice applications. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, 1993. This volume.

[70] D. Karis and K. M. Dobroth. Automating servies with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal of Selected Areas in Communications*, 9(4):574–585, 1991.

[71] H. Kautz. A circumscriptive theory of plan recognition. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. M.I.T. Press, Cambridge, Massachusetts, 1990.

[72] A. Kay and A. Goldberg. Personal dynamic media. *IEEE Computer*, 10(1):31–42, 1977.

[73] M. J. Kelly and A. Chapanis. Limited vocabulary natural language dialogue. *International Journal of Man-machine Studies*, 9:479–501, 1977.

[74] A. Kennedy, A. Wilkes, L. Elder, and W. S. Murray. Dialogue with machines. *Cognition*, 30(1):37–72, 1988.

[75] H. Kitano. ϕdm-dialog. *IEEE Computer*, 24(6):36–50, June 1991.

[76] R. M. Krauss and P. D. Bricker. Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, 41(2):286–292, 1967.

[77] R. M. Krauss and S. Weinheimer. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343–346, 1966.

[78] M. Kreuger. Responsive environments. In *Proceedings of the National Computer Conference*, 1977.

[79] F. Kubala, C. Barry, M. Bates, R. Bobrow, P. Fung, R. Ingria, J. Makhoul, L. Nguyen, R. Schwartz, and D. Stallard. BBN BYBLOS and HARC February 1992 ATIS benchmark results. In *Fifth DARPA Workshop on Speech and Natural Language*, San Mateo, Calif., 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.

[80] A. Kurematsu. Future perspective of automatic telephone interpretation. *Transactions of IEICE*, E75(1):14–19, January 1992.

[81] W. A Lea. Practical lessons from configuring voice I/O systems. In *Proceedings of Speech Tech/Voice Systems Worldwide*, New York, 1992. Media Dimensions, Inc.

[82] R. G. Leiser. Exploiting convergence to improve natural language understanding. *Interacting with Computers*, 1(3):284–298, December 1989.

[83] M. Lennig. Using speech recognition in the telephone network to automate collect and third-number-billed calls. In *Proceedings of Speech Tech'89*, pages 124–125, Arlington, Virginia, 1989. Media Dimensions.

[84] W. J. M. Levelt and S. Kelter. Surface form and memory in question-answering. *Cognitive Psychology*, 14(1):78–106, 1982.

[85] S. Levinson. Some pre-observations on the modelling of dialogue. *Discourse Processes*, 4(1), 1981.

[86] H. Levitt. Assistive speech technology. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, 1993. This volume.

[87] D. J. Litman and J. F. Allen. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200, 1987.

[88] D. J. Litman and J. F. Allen. Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 365–388. MIT Press, Cambridge, Massachusetts, 1990.

[89] P. A. Luce, T. C. Feustel, and D. B. Pisoni. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25(1):17–32, 1983.

[90] M. Marcus. New trends in natural language processing: The growth of statistical NLP. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, 1993. This volume.

[91] J. Mariani. Spoken language processing in the framework of human-machine communication at LIMSI. In *Proceedings of Speech and Natural Language Workshop*, pages 55–60, San Mateo, California, 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.

[92] R. E. Markinson. Associate clinical professor of surgery, University of California at San Francisco. Personal communication, June 1993.

[93] J. P. Marshall. A manufacturing application of voice recognition for assembly of aircraft wire harnesses. In *Proceedings of Speech Tech/Voice Systems Worldwide*, New York, 1992. Media Dimensions, Inc.

[94] G. L. Martin. The utility of speech input in user-computer interfaces. *International Journal of Man-machine Studies*, 30(4):355–375, 1989.

[95] T. B. Martin. Practical applications of voice input to machines. *Proceedings of the IEEE*, 64(4):487–501, April 1976.

[96] P. R. Michaelis, A. Chapanis, G. D. Weeks, and M. J. Kelly. Word usage in interactive dialog with restricted and unrestricted vocabularies. *IEEE Transactions on Professional Communication*, PC-20(4), December 1977.

[97] R. C. Moore. Integration of speech with natural language understanding. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, Washington, D. C., 1993.

[98] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth. Towards a reading coach that listens: Automated detection of oral reading errors. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)*, Menlo Park, California, 1993. American Association for Artificial Intelligence, AAAI Press/The MIT Press.

[99] I. R. Murray, J. L. Arnott, A. F. Newell, G. Cruickshank, K. E. P. Carter, and R. Dye. Experiments with a full-speed speech-driven word processor. Technical Report CS 91/09, Mathematics and Computer Science Department, University of Dundee, Dundee, Scotland, April 1991.

[100] C. Nakatani and J. Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 46–53, Columbus, Ohio, June 1993.

[101] R. Nakatsu. What does voice processing technology support today. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, 1993. This volume.

[102] A. F. Newell, J. L. Arnott, K. Carter, and G. Cruickshank. Listening typewriter simulation studies. *International Journal of Man-machine Studies*, 33(1):1–19, 1990.

[103] H. C. Nusbaum and E. C. Schwab. The effects of training on intelligibility of synthetic speech: II. The learning curve for synthetic speech. In *Proceedings of the 105th meeting of the Acoustical Society of America*, Cincinnati, Ohio, May 1983.

[104] J. M. Nye. Human factors analysis of speech recognition systems. In *Speech Technology I*, pages 50–57, 1982.

[105] R. B. Ochsman and A. Chapanis. The effects of 10 communication modes on the behaviour of teams during co-operative problem-solving. *International Journal of Man-Machine Studies*, 6(5):579–620, Sept. 1974.

[106] Committee on Computerized Speech Recognition Technologies. *Automatic Speech Recognition in Severe Environments.* Commission on Engineering and Technical Systems, National Research Council, National Academy of Sciences Press, Washington, D. C., 1984.

[107] S. L. Oviatt. Pen/voice: Complementary multimodal communication. In *Proceedings of Speech Tech'92*, pages 238–241, New York, February 1992.

[108] S. L. Oviatt. Predicting spoken disfluencies during human-computer interaction. In K. Shirai, editor, *Proceedings of the International Symposium on Spoken Dialogue: New Directions in Human-Machine Communication*, Tokyo, Japan, November 1993.

[109] S. L. Oviatt. Toward multimodal support for interpreted telephone dialogues. In M. M. Taylor, F. Néel, and D. G. Bouwhuis, editors, *Structure of Multimodal Dialogue.* Elsevier Science Publishers B. V., Amsterdam, Netherlands, in press.

[110] S. L. Oviatt and P. R. Cohen. The contributing influence of speech and interaction on human discourse patterns. In J. W. Sullivan and S. W. Tyler, editors, *Intelligent User Interfaces*, chapter 3, pages 69–83. ACM Press Frontier Series, Addison-Wesley Publishing Co., New York, New York, 1991.

[111] S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 5(4):297–326, 1991a.

[112] S. L. Oviatt, P. R. Cohen, M. W. Fong, and M. P. Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In J. Ohala, editor, *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 1351–1354, University of Alberta, October 1992.

[113] S. L. Oviatt, P. R. Cohen, M. Wang, and J. Gaston. A simulation-based research strategy for designing complex NL systems. In *ARPA Human Language Technology Workshop*, Princeton, New Jersey, March 1993.

[114] D. S. Pallett, J. G. Fiscus, W. M. Fisher, and J. S. Garofolo. Benchmark tests for the DARPA spoken language program. In *Proc. of the ARPA Workshop on Human Language Technology*, San Mateo, Calif., 1993. Advanced Research Projects Agency, Morgan Kaufmann, Publishers, Inc.

[115] S. Pavan and B. Pelletti. An experimental approach to the design of an oral cooperative dialogue. In *Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218).* Commission of the European Communities, 1990.

[116] J. Peckham. Speech understanding and dialogue over the telephone: An overview of the ESPRIT SUNDIAL project. In *Proc. of the Speech and Natural Language Workshop*, pages 14–28, San Mateo, California, February 1991. DARPA/ISTO, Morgan Kaufmann Publishers, Inc.

[117] C. R. Perrault and J. F. Allen. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics*, 6(3):167–182, 1980.

[118] E. Petajan, B. Bradford, D. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. In *Proc. of Human Factors in Computing Systems (CHI'88)*, pages 19–25, New York, 1988. Association for Computing Machinery, ACM Press.

[119] R. Polany and R. Scha. A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 413–419, Stanford, California, 1984.

[120] A. Pollack. Computer translator phones try to compensate for Babel. *New York Times*, January 29, 1993.

[121] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. In *Proceedings of the Speech and Natural Language Workshop*, pages 372–377, San Mateo, California, October 1991. Morgan Kaufmann, Publishers, Inc.

[122] P. J. Price. Evaluation of spoken language systems: The ATIS domain. In *Proceedings of the 3rd DARPA Workshop on Speech and Natural Language*, pages 91–95, San Mateo, California, 1990. Morgan Kaufmann Publishers, Inc.

[123] L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg. A voice-controlled, repertory-dialer system. *Bell System Technical Journal*, 59(7):1153–1163, September 1980.

[124] H. Rheingold. *Virtual Reality*. Summit Books, 1991.

[125] D. B. Roe, F. Pereira, R. W. Sproat, and M. D. Riley. Toward a spoken language translator for restricted-domain context-free languages. In *Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology*, pages 1063–1066, Genova, Italy, 1991. European Speech Communication Association.

[126] F. A. Rosenhoover, J. S. Eckel, F. A. Gorg, and S. W. Rabeler. AFTI/F-16 voice interactive avionics evaluation. In *Proceedings of National Aerospace and Electronics Conference (NAECON'87)*. IEEE, 1987.

[127] J. Rubin-Spitz and D. Yashchin. Effects of dialogue design on customer responses in automated operator services. In *Proceedings of Speech Tech'89*, 1989.

[128] A. I. Rudnicky. Mode preference in a simple data-retrieval task. In *ARPA Human Language Technology Workshop*, Princeton, New Jersey, March 1993.

[129] J. R. Searle. *Speech acts: An essay in the philosophy of language.* Cambridge University Press, Cambridge, 1969.

[130] B. Shneiderman. Natural vs. precise concise languages for human operation of computers: Research issues and experimental approaches. In *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics*, pages 139–141, Philadelphia, Pennsylvania, June 1980.

[131] B. Shneiderman. *Software Psychology: Human Factors in Computer and Information systems.* Winthrop Publishers, Inc., Cambridge, Massachusetts, 1980.

[132] B. Shneiderman. Direct manipulation: A step beyond programming languages. *IEEE Computer*, 16(8):57–69, 1983.

[133] E. Shriberg, E. Wade, and P. Price. Human-machine problem-solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proceedings of Speech and Natural Language Workshop*, pages 49–54, San Mateo, California, 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.

[134] C. Sidner and D. Israel. Recognizing intended meaning and speaker's plans. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 203–208, Vancouver, B. C., 1981.

[135] C. A. Simpson, C. R. Coler, and E. M. Huff. Human factors of voice I/O for aircraft cockpit controls and displays. In *Proceedings of the Workshop on Standardization for Speech I/O Technology*, pages 159–166, Gaithersburg, Maryland, March 1982. National Bureau of Standards.

[136] C. A. Simpson, M. E. McCauley, E. F. Roland, J. C. Ruth, and B. H. Williges. System design for speech recognition and generation. *Human Factors*, 27(2):115–141, 1985.

[137] C. A. Simpson and T. N. Navarro. Intelligibility of computer generated speech as a function of multiple factors. In *Proceedings of the National Aerospace and Electronics Conference (NAECON)*, pages 932–940, New York, May 1984. IEEE.

[138] D. Small and L. Weldon. An experimental comparison of natural and structured query languages. *Human Factors*, 25:253–263, 1983.

[139] J. Spitz. Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, Asilomar, California, February 1991. Defense Advanced Research Projects Agency.

[140] D. Stallard and R. Bobrow. Fragment processing in the DELPHI system. In *Proceedings of the Speech and Natural Language Workshop*, pages 305–310, San Mateo, Calif., Feb. 1992. DARPA, Morgan Kaufmann.

[141] R. L. Jr. Street, R. M. Brady, and W. B. Putman. The influence of speech rate stereotypes and rate similarity on listeners' evaluations of speakers. *Journal of Language and Social Psychology*, 2(1):37–56, 1983.

[142] L. A. Streeter, D. Vitello, and S. A. Wonsiewicz. How to tell people where to go: comparing navigational aids. *International Journal of Man-Machine Studies*, 22:549–562, 1985.

[143] R. F. Swider. Operational evaluation of voice command/response in an Army helicopter. In *Proceedings of Military Speech Tech 1987*, pages 143–146, Arlington, Virginia, 1987. Media Dimensions.

[144] S. Tanaka, D. K. Wild, P. J. Seligman, W. E. Halperin, V. Behrens, and V. Putz-Anderson. Prevalence and work-relatedness of self-reported carpal tunnel syndrome among U.S. workers — analysis of the Occupational Health Supplement Data of 1988 National Health Interview Survey. National Institute of Occupational Safety and Health, and Centers for Disease Control and Prevention (Cincinnati), in submission.

[145] H. R. Tenant, K. M. Ross, R. M. Saenz, C. W. Thompson, and J. R. Miller. Menu-based natural language understanding. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 151–158, Cambridge, Massachusetts, June 1983.

[146] J. C. Thomas, M. B. Rosson, and M. Chodorow. Human factors and synthetic speech. In B. Shackel, editor, *Proceedings of INTERACT'84*, Amsterdam, 1984. Elsevier Science Publishers B. V. (North Holland).

[147] J. A. Turner, M. Jarke, E. A. Stohr, Y. Vassiliou, and N. White. Using restricted natural language for data retrieval: A plan for field evaluation. In Y. Vassiliou, editor, *Human Factors and Interactive Computer systems*, chapter 8, pages 163–190. Ahlex Publishing Corp., Norwood, N. J., 1984.

[148] A. F. VanKatwijk, F. L. VanNes, H. C. Bunt, H. F. Muller, and F. F. Leopold. Naive subjects interacting with a conversing information system. *IPO Annual Progress Report*, 14:105–112, 1979.

[149] D. Visick, P. Johnson, and J. Long. The use of simple speech recognisers in industrial applications. In *Proceedings of INTERACT'84 First IFIP Conference on Human-Computer Interaction*, London, U.K., 1984.

[150] J. W. Voorhees, N. M. Bucher, E. M. Huff, C. A. Simpson, and D. H. Williams. Voice interactive electronic warning system (views). In *Proceedings of the IEEE/AIAA 5th Digital Avionics Systems Conference*, pages 3.5.1–3.5.8, New York, 1983. IEEE.

[151] W. Wahlster. User and discourse models for multimodal communication. In J. W. Sullivan and S. W. Tyler, editors, *Intelligent User Interfaces*, chapter 3, pages 45–68. ACM Press Frontier Series, Addison Wesley Publishing Co., New York, New York, 1991.

[152] C. Weinstein. Opportunities for advanced speech processing in military computer-based systems. *Proceedings of the IEEE*, 79(11):1626–1641, November 1991.

[153] C. J. Weinstein. Military/government applications of speech processing technology. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, Washington, D. C., 1993.

[154] J. Welkowitz, S. Feldstein, M. Finkelstein, and L. Aylesworth. Changes in vocal intensity as a function of interspeaker influence. *Perceptual and Motor Skills*, 10:715–718, 1972.

[155] J. T. Williamson. Flight test results of the AFTI/F-16 voice interactive avionics program. In *Proceedings of the American Voice I/O Society (AVIOS) 87 Voice I/O Systems Applications Conference*, pages 335–345, Alexandria, Virginia, 1987.

[156] J. G. Wilpon. Applications of voice processing in telecommunications. In D. B. Roe and J. Wilpon, editors, *Human-Machine Communication by Voice*. National Academy of Sciences Press, 1993. This volume.

[157] T. Winograd. *Understanding Natural Language*. Academic Press, New York, 1972.

[158] T. Winograd and F. Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Ablex Publishing Co., Norwood, New Jersey, 1986.

[159] T. Yamaoka and H. Iida. Dialogue interpretation model and its application to next utterance prediction for spoken language processing. In *Proceedings of Eurospeech'91: 2nd European Conference on Speech Communication and Technology*, pages 849–852, Genova, Italy, 1991. European Speech Communication Association.

[160] F. Yato, T. Takezawa, S. Sagayama, J. Takami, H. Singer, N. Uratani, T. Morimoto, and A. Kurematsu. International joint experiment toward interpreting telephony (in Japanese). Technical report, The Institute of Electronics, Information, and Communication Engineers, 1992.

[161] 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), February 1989.

[162] E. Zoltan-Ford. *Language Shaping and Modeling in Natural Language Interactions with Computers*. PhD thesis, Johns Hopkins University, Baltimore, Maryland, 1983. Psychology Department.

[163] E. Zoltan-Ford. Reducing variability in natural-language interactions with computers. In M. J. Alluisi, S. de Groot, and E. A. Alluisi, editors, *Proceedings of the Human Factors Society — 28th Annual Meeting*, volume 2, pages 768–772, San Antonio, Texas, 1984.

[164] E. Zoltan-Ford. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527–547, 1991.

[165] V. Zue, J. Glass, D. Goddeau, D. Goodine, L. Hirschman, M. Phillips, J. Polifroni, and S. Seneff. The MIT ATIS system: February 1992 progress report. In *Fifth DARPA Workshop on Speech and Natural Language*, San Mateo, Calif., 1992. Defense Advanced Research Projects Agency, Morgan Kaufmann Publishers, Inc.