Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9921–9927, October 1995
Colloquium Paper

This paper is a condensed version of one that was present at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irving, CA, February 8–9, 1993.

The role of voice input for human–machine communication

PHILIP R. COHEN AND SHARON L. OVIATT

Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, OR 97034

ABSTRACT **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 with vocabularies of 1000 to 2000 words, and larger. Already, computer manufacturers are building speech recognition subsystems into their new product lines. However, before this technology can be broadly useful, a substantial knowledge base is needed about human spoken language and performance during computer-based spoken interaction. This paper reviews application areas in which spoken interaction can play a significant role, assesses potential benefits of spoken interaction with machines, and compares voice with other modalities of human–computer interaction. It also discusses information that will be needed to build a firm empirical foundation for the design of future spoken and multimodal interfaces. Finally, it argues for a more systematic and scientific approach to investigating spoken input and performance with future language technology.**

From the beginning of the computer era, futurists have dreamed of the conversational computer—a machine that we could engage in natural spoken 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. Despite prolonged research and many notable scientific and technological achievements, there have been few real human–computer dialogues until recently, and those existing have been keyboard exchanges rather than spoken. This situation has begun to change, however. Steady progress in speech recognition and natural language processing technologies, supported by dramatic advances in computer hardware, has enabled laboratory prototype systems with which one can conduct simple question-answering dialogues. Although far from human-level conversation, this initial capability is generating considerable optimism for the future of human–computer interaction using voice.

This paper aims to identify applications for which spoken interaction is advantageous, to clarify the role of voice with respect to other modalities of human–computer interaction, and to consider obstacles to the successful development and commercialization of spoken language systems.

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.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

WHEN IS SPEAKING TO 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 the section *Natural Language Interaction.*

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. 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 (1). 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.

Other hands/eyes-busy applications that have benefited from voice interaction include data entry and machine control in factories and field applications (2), access to information for military command-and-control, cockpit management (3, 4), astronauts' information management during extra-vehicular access in space, dictation of medical diagnoses, maintenance and repair

of equipment, control of automobile equipment (e.g., radios, telephones, climate control), and navigational aids.

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. Moreover, results showing the elimination of benefits once error correction is considered have been found in tasks as simple as entry of connected digits (5). However, it is claimed that most failures of speech technology have been the result of poor human factors engineering and management (6), rather than low recognition accuracy per se.

Limited Keyboard/Screen Option. The most prevalent current use of speech recognition are telephone-based applications that replace or augment operator services (e.g., collect calls), handling hundreds of millions of callers each year and resulting in multi-million dollar savings (7–9). 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.*

The considerable success at automating the simpler operator services opens the possibility for more ambitious telephone-based applications, such as information access from remote data bases. For example, the caller might inquire about airline and train schedules (11, 12), yellow-pages information, or bank account balances (8), 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 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" could improve such tasks.

Perhaps the most challenging potential application of telephone-based spoken language technology is the interpretation of telephony (13, 14) in which two callers speaking different languages can engage in a dialogue mediated by a spoken language translation system. Such systems are currently designed to incorporate speech recognition, machine translation, and speech synthesis subsystems, and to interpret one sentence at a time.

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 (15, 16), 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.

Disability. A major potential use of voice technology will be to assist deaf users in communicating with the hearing world using a telephone (17). Speech recognition could also be used by motorically impaired users to control suitably augmented household appliances, wheelchairs, and robotic prostheses. 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 ≈1.5% of office workers in occupations that typically involve the use of keyboards (18). However, speech recognizers may themselves lead to different repetitive stress injuries (19).

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 (20, 21). 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 the speech recognition problem for such applications may be simplified because the words being spoken are supplied by the computer, the recognition system nonetheless will be confronted with mispronunciations and altered articulation, requiring a degree of robustness not often considered in other applications of speech recognition.

Summary

There are numerous existing applications of voice-based human–computer interaction, and new opportunities are emerging rapidly. In many applications for which the user's input can be constrained sufficiently to permit high recognition accuracy, voice input has led to faster task performance and fewer errors than keyboard entry. Unfortunately, no reliable method yet exists to predict when voice input will be the most effective, efficient, or preferred modality of communication.

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.

COMPARISON OF SPOKEN LANGUAGE WITH OTHER MODES OF COMMUNICATING

A user speaking to a machine typically expects to be able to speak in natural language; that is, to use ordinary linguistic constructs delivered in a conversational manner. Conversely, if natural language interaction is chosen as a modality of human–computer communication, users often 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 other modalities.

Spoken Language System Prototypes

Research is progressing toward 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 (11), developed at Bolt, Beranek and Newman (22), Carnegie–Mellon University (23), Massachusetts Institute of Technology (24), SRI International (25), and other institutions, allow novice users to obtain information in real-time from the Official Airline Guide data base, through 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 the Official Airline Guide's data base, and produce a tabular set of answers that satisfy the

*An excellent review of the human factors and technical difficulties encountered in telecommunications applications of speech recognition can be found in Karis and Dobroth (10).

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 also are ongoing in Europe (12, 26) and Japan (27).

Comparison of Language-Based Communication Modalities

In a series of studies of interactive human–human communication, Chapanis and colleagues (28–32) compared the efficiency of human–human communication when subjects used any of 10 communication modalities, including face-to-face, voice-only, linked teletypes, and interactive handwriting. The most important determinant of a team's problem-solving speed was reported 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 Fig. 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 (33, 34), finding a 3-fold speed advantage for interactive telephone speech over keyboard communication, as well as differences in dialogue structure. In a study comparing voice and handwritten interaction with a simulated computer system (35), voice input resulted in 50–100% faster tasks completion. It also contained more words, a more variable vocabulary, and more syntactic ambiguity than handwritten input.

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 communication modalities are options. Whereas some studies report a decided preference for speech in comparison with other modalities (36), other studies report the opposite conclusion (37, 38). Thus, despite the potential benefits of human–computer interaction using voice, it is not obvious why people should want to speak to their computers in performing many tasks—in particular, 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.

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,

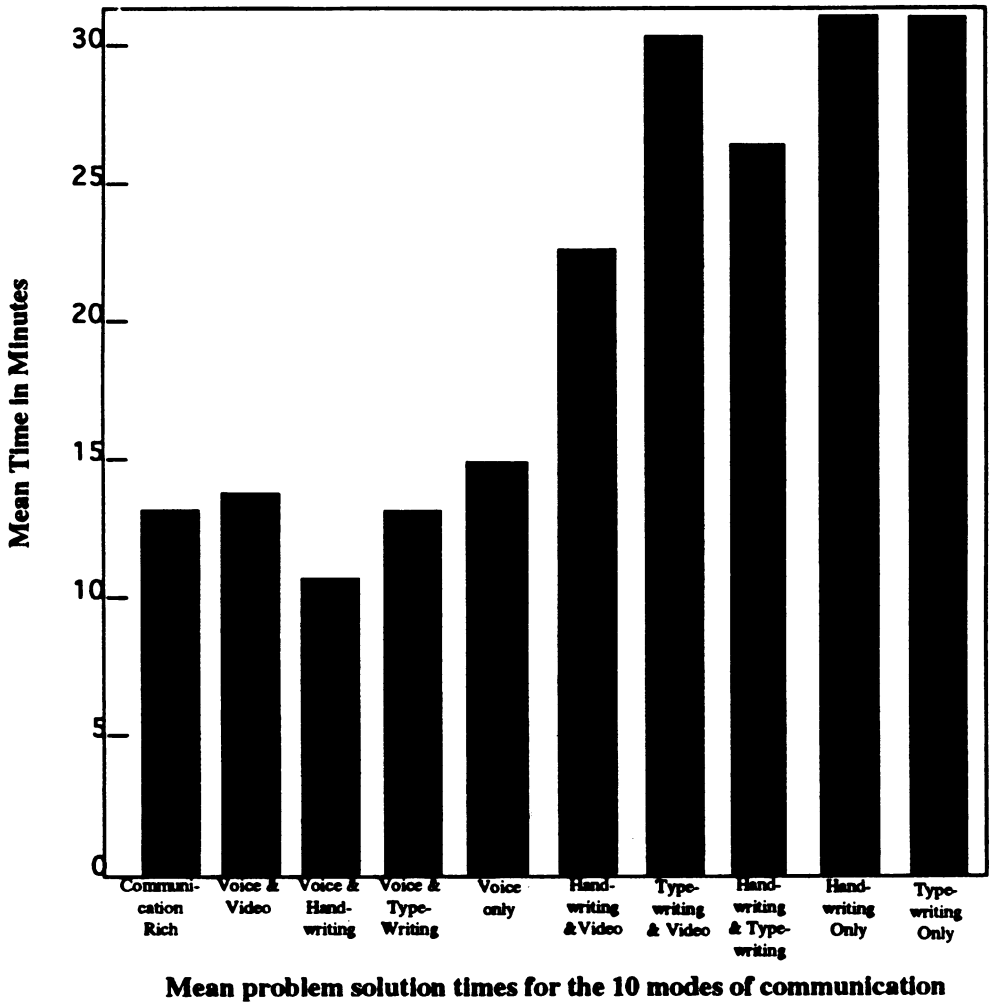

FIG. 1. Voice determines task efficiency (32).

pointing and gesturing with devices such as the mouse, a digitizing pen, trackballs, touchscreens, and digitizing gloves. It is important to understand what role spoken language can play in supporting human computer interaction.

Graphical User Interfaces and Direct Manipulation. The graphical user-interface (GUI) paradigm employs techniques pioneered at SRI International and at Xerox's Palo Alto Research Center in the late 1960s and 1970s (39, 40). This paradigm, which later was popularized by the Apple Macintosh and by Microsoft Windows, offers the user menus, icons, and pointing devices, such as the "mouse" (41), as well as multiple windows in which to display output. With GUIs, users perform actions by selecting objects and then choosing the desired action from a menu, rather than by typing commands.

With many GUIs, a user can directly manipulate graphical objects in order to perform actions on the objects they 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. 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, users now can directly manipulate 3-dimensional virtual worlds using computer-instrumented gloves and body-suites (42), permitting body motion to affect the virtual environment.

Strengths. Many writers have identified virtues of graphically-based direct manipulation interfaces or DMIs (43, 44), 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. DMIs do not suffice for all needs, however. One weakness is the paucity of means for identifying entities. Merely allowing users to select currently displayed entities provides them little support, beyond simple string matching, for identifying objects not on the screen, for specifying temporal relations that denote future or past events, for identifying and operating on large sets of entities, or for exploiting the context of interaction. What is missing in DMIs 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.†

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 their menu location. Even when the user knows the location of the desired command, navigating the hierarchy still requires time and effort.

Because direct manipulation emphasizes rapid, graphical response to actions, the time at which an action occurs is literally the time at which it 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 some unknown but describable future time.

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

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 studies of a direct-manipulation VLSI design system augmented with speaker-dependent speech recognition indicate that users are as fast at speaking single-word commands as they are at invoking the same commands with mouse-button clicks, or by typing a single letter command abbreviation (45). Furthermore, circuit designers were able to complete 24% more tasks when spoken commands were available than when they used only a keyboard and mouse interface. In a recent study of human–computer interaction to retrieve information from a small data base 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 (36). 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 task accuracy.

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, and other natural languages, 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 descriptions are deployed simultaneously, in referring to the sentence subject and object(s), and in describing some event in which those entities are participating. Furthermore, natural language commands can shortcut the navigation of a menu hierarchy to invoke known commands.

Ideally, natural language systems require only a minimum of training on the system domain. The system should have sufficient vocabulary, as well as linguistic, semantic, and dialogue capabilities, to support interactive problem solving by infrequent users. For example, at its present state of development, many users can successfully solve travel-planning problems with one of the ATIS systems (11) within a few minutes of introduction to the system and its coverage. To develop systems with this level of robustness using present statistical language-modeling techniques, the system must be trained and tested on a substantial amount of data representing input from a broad spectrum of users.‡ Currently, the level of training required to achieve a given level of proficiency in using these systems is unknown.

Weaknesses. Various disadvantages are apparent when natural language is incorporated into an interface. Pure natural language systems tend to suffer from opaque linguistic and conceptual coverage. That is, the user knows the system cannot interpret every utterance, but does not know precisely what it *can* interpret (46, 47). 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 (48), lead to frustration and disillusionment.

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.

Another disadvantage 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 presenta-

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

tions 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.

Summary: Circumstances Favoring Spoken Language Interaction with Machines

Empirical results on the circumstances favoring voice-based interaction, as well as an analysis of interactions for which natural language may be most appropriate, indicates 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 data base, or creating rules for action (e.g., "If I am late for a meeting, notify the meeting participants").

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 (49). Multimodal systems are discussed further below. However, before spoken language systems can be deployed on a wide scale, numerous obstacles need to be overcome.

RESEARCH DIRECTIONS FOR SPOKEN LANGUAGE SYSTEMS

Although there are numerous technical challenges to building spoken language systems, many of which are detailed in this volume, much further research is needed to build usable systems that incorporate spoken language. Below, we consider information needed about spontaneous speech, spoken natural language, spoken dialogue, and multimodal interaction.

Spontaneous Speech

When an utterance is spontaneously spoken, it may well involve 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 present technology can be deployed robustly. Current research has begun to investigate techniques for detecting and handling disfluencies in spoken human–computer interaction (50–52). Alternatively, user interface techniques have been developed that can minimize the number of disfluencies that occur (53), based on observed relationships between the rate of disfluencies and both utterance length and degree of structure in the system's presentation format.

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. 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, for which the 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 levels of recognition and understanding are needed for them to solve problems effectively
- How much and what kind of training may be 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 (54). If users can discern the system's vocabulary, one 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 (30, 31). User adaption to vocabulary restrictions has also been found for simulated human–computer interaction (55), 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. Numerous studies of human communication have shown that people will adopt the speech styles of their interlocutors [see Giles *et al.* (56) for a survey]. However, it is not known if the modeling of syntactic structures occurs in *spoken* human–computer interaction. 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 expected words and phrases (57). 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 (35). Highly structured spoken interactions can reduce the perplexity and syntactic ambiguity of the user's speech by more than 70%, thereby simplifying the system's language processing. At the same time, for service-oriented tasks, research has shown that users sometimes prefer structured spoken interaction over unconstrained ones by as much as a factor of 2-to-1 (35).

Interaction and Dialogue

Present spoken language systems have supported question-answer dialogues (11), or dialogues in which the user is prompted for information (12, 58). To support a broader range of dialogue behavior, more general models of dialogue are being investigated, both mathematically and computationally. These include both dialogue grammars and plan-based models of dialogue. The dialogue grammar approach models dialogue simply as a finite state transition network (59, 60), 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 (58, 61). However, a number of drawbacks to the model are evident (62, 63). First, it requires that the communicative action(s) performed by the speaker in issuing an utterance be identified. Second, 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 an appropriate role as a cooperative conversant. Some functional equivalent of planning is thus likely to be required.

Plan-based models (64, 65) 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 (66), such as requesting, informing, warning, suggesting, and confirming. These models propose that the listener's job is to

uncover and respond appropriately to the speaker's underlying plan, rather than just to the utterance. Current research guided by this model is attempting to incorporate more complex dialogue phenomena, such as clarifications (67, 68), and to model dialogue more as a joint enterprise between the participants (69–71, 78).

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 would count as acceptable dialogue behavior. As in other branches of computer science, such specifications may lead to methods for mathematically and empirically evaluating whether a given system has met the specifications. 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.

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 probably will be multimodal, in the sense of combining voice with screen feedback, use of a pointing device, gesturing, handwriting, and other modalities (16, 72, 73).

Among the many advantages of multimodal systems are the following:

Enhanced Error Avoidance and Correction. Multimodal interfaces offer the opportunity for users to avoid errors that would otherwise occur in a unimodal interface (16). For example, users can select to write rather than speak a difficult-to-pronounce foreign surname. Furthermore, when repeat "spiral" errors are encountered during spoken input, an alternate mode enables shortcutting them (16, 74). Multimodal systems also have the potential to increase the recognition rate in adverse environments through fusion of multiple sources of information (75, 76).

Accommodation of Various Situations, Users, and Task. A change in the environment of portable computer use may alter people's preferences to employ one modality of communication over another. For example, public environments that are noisy, or in which privacy is an issue, often are ones in which people prefer not to speak. Likewise, individual and task differences can strongly influence people's willingness to use one input mode over another. Multimodal systems thus have the potential to accommodate a wider array of different users, tasks, and situations than unimodal ones (16).

User Preference. Users may strongly prefer multimodal interaction. In a recent comparison of spoken, written, and combined pen/voice input, it was found that 56–89% of users preferred interacting multimodally, which was perceived to be easier and more flexible (73).

SCIENTIFIC RESEARCH ON SPOKEN AND MULTIMODAL INTERACTION WITH COMPUTERS

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 (77). 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 real environments rather than the laboratory. There is little research on an interaction's dependence on the modality used, or on the task, 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 beginning to be addressed.

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

- Longitudinal studies of users' linguistic and problem-solving behavior to examine how users adapt over time to speech input 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 channeling user input to match system capabilities
- Studies comparing the relative effectiveness and usability of spoken language technology with other input alternatives
- Studies analyzing users' language, task performance, and preferences to use different modalities, alone and within an integrated multimodal interface

The information gained from such studies would be an invaluable addition to our knowledge of how spoken language processing can best be woven into a robust and usable human–computer interface. Additional research will be needed to understand how to build limited but robust dialogue systems based on a variety of communication modalities, and to improve our basic understanding of the nature of dialogue. Finally, an important but underappreciated requirement to the successful deployment of spoken language technology is the development of empirically-validated guidelines for creating interfaces that incorporate spoken language. Such guidelines can inform developers in advance about the tradeoffs associated with different interface design decisions, and their likelihood of yielding a more optimal system.

Many thanks to Jared Bernstein, Clay Coler, Carol Simpson, Ray Perrault, Robert Markinson, Raja Rajasekharan, and John Vester for valuable discussions and source materials. The writing of this paper was supported in part by Grant IRI-9213472 from the National Science Foundation.

1. Marshall, J. P. (1992) in *Proceedings of Speech Tech/Voice Systems Worldwide* (Media Dimensions, New York).
2. Martin, T. B. (1976) *Proc. IEEE* **64,** 487–501.
3. Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C. & Williges, B. H. (1985) *Hum. Factors* **27,** 115–141.
4. Weinstein, C. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10011–10016.
5. Hauptmann, A. G. & Rudnicky, A. I. (1990) in *Proceedings of the Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA), pp. 219–224.
6. Lea, W. A. (1992) in *Proceedings of Speech Tech/Voice Systems Worldwide* (Media Dimensions, New York).
7. Lennig. M. (1989) in *Proceedings of Speech Tech '89* (Media Dimensions, Arlington, VA), pp. 124–125.
8. Nakatsu, R. & Suzuki, Y. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10023–10030.
9. Wilpon, J. G. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9991–9998.
10. Karis, D. & Dobroth, K. M. (1991) *IEEE J. Selected Areas Commun.* **9,** 574–585.
11. Advanced Research Projects Agency (1993) *ARPA Spoken Language Systems Technology Workshop* (Massachusetts Institute of Technology, Cambridge).
12. Peckham, J. (1991) in *Proceedings of the Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA), pp. 14–28.
13. Kurematsu, A. (1991) *Trans. IEICE* **E75,** 14–19.
14. Roe, D. B., Pereira, F., Sproat, R. W. & Riley, M. D. (1991) in *Proceedings of Eurospeech '91: Second European Conference on Speech Communication and Technology* (Eur. Speech Commun. Assoc., Genoa, Italy), pp. 1063–1066.
15. Crane, H. D. (1991) *Business Intelligence Program Report D91-1557* (SRI International, Menlo Park, CA).

16. Oviatt, S. L. (1992) in *Proceedings of Speech Tech '92* (Media Dimensions, New York), pp. 238–241.
17. Bernstein, J. (1989) in *Speech to Text: Today and Tomorrow*, eds. Harkins, J. E. & Virvan, B. M. (Gallaudet Univ. Res. Institute, Washington, DC), GRI Monograph Series B., No. 2.
18. Tanaka, S., Wild, D. K., Seligman, P. J., Halperin, W. E., Behrens, V. & Putz-Anderson, V., in *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 press.
19. Markinson, R. E. (1993) Personal communication (Univ. of California, San Francisco).
20. Bernstein, J., Cohen, M., Murveit, H., Rtischev, D. & Weintraub, M. (1990) in *Proceedings of the 1990 International Conference on Spoken Language Processing* (Acoustical Soc. of Japan, Kobe), pp. 1185–1188.
21. Mostow, J., Hauptmann, A. G., Chase, L. L. & Roth, S. (1993) in *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)* (AAAI/MIT Press, Menlo Park, CA).
22. Kubala, F., Barry, C., Bates, M., Bobrow, R., Fung, P., Ingria, R., Makhoul, J., Nguyen, L., Schwartz, R. & Stallard, D. (1992) in *Fifth DARPA Workshop on Speech and Natural Language* (Kaufmann, San Mateo, CA).
23. Huang, X., Alleva, F., Hwang, M.-Y. & Rosenfeld, R. (1993) in *Proceedings of the ARPA Workshop on Human Language Technology* (Kaufmann, San Mateo, CA).
24. Zue, V., Glass, J., Goddeau, D., Goodine, D., Hirschman, L., Phillips, M., Polifroni, J. & Seneff, S. (1992) in *Fifth DARPA Workshop on Speech and Natural Language* (Kaufmann, San Mateo, CA).
25. Appelt, D. E. & Jackson, E. (1992) in *Fifth DARPA Workshop on Speech and Natural Language* (Kaufmann, San Mateo, CA).
26. Mariani, J. (1992) in *Proceedings of Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA), pp. 55–60.
27. Yato, F., Takezawa, T., Sagayama, S., Takami, J., Singer, H., Uratani, N., Morimoto, T. & Kurematsu, A. (1992) *Technical Report* (The Institute of Electronics, Information, and Communication Engineers, Tokyo).
28. Chapanis, A., Ochsman, R. B., Parrish, R. N. & Weeks, G. D. (1972) *Hum. Factors* **14,** 487–509.
29. Chapanis, A., Ochsman, R. B., Parrish, R. N. & Weeks, G. D. (1977) *Hum. Factors* **19,** 101–125.
30. Kelly, M. J. & Chapanis, A. (1977) *Int. J. Man–Mach. Stud.* **9,** 479–501.
31. Michaelis, P. R., Chapanis, A., Weeks, G. D. & Kelly, M. J. (1977) *IEEE Trans. Prof. Commun.* **PC-20.**
32. Ochsman, R. B. & Chapanis, A. (1974) *Int. J. Man–Mach. Stud.* **6,** 579–620.
33. Cohen, P. R. (1984) *Computat. Linguist.* **10,** 97–146.
34. Oviatt, S. L. & Cohen P. R. (1991) in *Intelligent User Interfaces*, ACM Press Frontier Series, eds. Sullivan J. W. & Tyler, S. W. (Addison-Wesley, New York), pp. 69–83.
35. Oviatt, S. L., Cohen, P. R. & Wang, M. Q. (1994) *Speech Commun.* **15,** 283–300.
36. Rudnicky, A. I. (1993) in *ARPA Human Language Technology Workshop* (Princeton, NJ).
37. Murray, I. R., Arnott, J. L., Newell, A. F., Cruickshank, G., Carter, K. E. P. & Dye, R. (1991) *Technical Report CS 91/09* (Mathematics and Computer Science Department, Univ. of Dundee, Dundee, Scotland).
38. Newell, A. F., Arnott, J. L., Carter, K. & Cruickshank, G. (1990) *Int. J. Man–Mach. Stud.* **33,** 1–19.
39. Englebart, D. (1973) in *National Computer Conference*, pp. 221–227.
40. Kay, A. & Goldberg, A. (1977) *IEEE Comput.* **10,** 31–42.
41. English, W. K., Englebart, D. C. & Berman, M. A. (1967) *IEEE Trans. Hum. Factors Electron.* **HFE-8,** 5–15.
42. Rheingold, H. (1991) *Virtual Reality* (Summit Books, New York).
43. Hutchins, E. L., Hollan, J. D. & Norman, D. A. (1986) in *User Centered System Design*, eds. Norman, D. A. & Draper, S. W. (Erlbaum, Hillsdale, NJ), pp. 87–124.
44. Shneiderman, B. (1983) *IEEE Comput.* **16,** 57–69.
45. Martin, G. L. (1987) *Int. J. Man–Mach. Stud.* **30,** 355–375.
46. Small, D. & Weldon, L. (1983) *Hum. Factors* **25,** 253–263.
47. Turner, J. A., Jarke, M., Stohr, E. A., Vassiliou, Y. & White, N. (1984) in *Human Factors and Interactive Computer Systems*, ed. Vassiliou, Y. (Ablex, Norwood, NJ), pp. 163–190.
48. Shneiderman, B. (1992) *Designing the User Interface: Strategies for Effective Human–Computer Interaction* (Addision-Wesley, Reading, MA).
49. Cohen, P. R. (1992) in *Proceedings of UIST'92* (ACM, New York), pp. 143–149.
50. Bear, J., Dowding, J. & Shriberg, E. (1992) in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* (Newark, DE).
51. Hindle, D. (1983) in *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics* (Cambridge, MA), pp. 123–128.
52. Nakatani, C. & Hirschberg, J. (1993) in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (Columbus, OH), pp. 46–53.
53. Oviatt, S. L. (1995) *Comput. Speech Lang.* **9,** 19–35.
54. Cole, R., Hirschman, L., Atlas, L., Beckman, M., Bierman, A., Bush, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorain, S. & Zue, V. (1992) *Technical Report CS/E 92-014* (Oregon Graduate Institute, Portland).
55. Zoltan-Ford, E. (1991) *Int. J. Man–Mach. Stud.* **34,** 527–547.
56. Giles, H., Mulac, A., Bradac, J. J. & Johnson, P. (1987) in *Communication Yearbook 10*, ed. McLaughlin, M. L. (Sage, Beverly Hills, CA), pp. 13–48.
57. Basson, S. (1992) in *Proceedings of COST232 Workshop—European Cooperation in Science and Technology.*
58. Andry, F. (1992) in *Proceedings of the International Conference on Spoken Language Processing* (Banff, AB, Canada).
59. Dahlbäck, N. & Jönsson, A. (1992) in *Proceedings of the 14th Annual Conference of the Cognitive Science Society (COGSCI-92)* (Bloomington, IN).
60. Winograd, T. & Flores, F. (1986) *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ).
61. Young, S. R., Hauptmann, A. G., Ward, W. H., Smith, E. T. & Werner, P. (1989) *Commun. ACM* **32.**
62. Levinson, S. (1981) *Discourse Processes* **4.**
63. Cohen, P. R. (1994) in *Cognitive Processing for Vision and Voice: Proceedings of the Fourth NEC Research Symposium.*
64. Cohen, P. R. & Perrault, C. R. (1972) *Cognit. Sci.* **3,** 177–212.
65. Perrault, C. R. & Allen, J. F. (1980) *Am. J. Comput. Linguist.* **6,** 167–182.
66. Searle, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language* (Cambridge Univ. Press, Cambridge, U.K.).
67. Grosz, B. & Sidner, C. (1986) *Comput. Linguist.* **12,** 175–204.
68. Litman, D. J. & Allen, J. F. (1987) *Cognit. Sci.* **11,** 163–200.
69. Clark, H. H. & Wilkes-Gibbs, D. (1986) *Cognition* **22,** 1–39.
70. Cohen, P. R. & Levesque, H. J. (1991) in *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (Kaufmann, San Mateo, CA), pp. 951–957.
71. Grosz, B. & Sidner, C. (1990) in *Intentions in Communication*, eds. Cohen, P. R., Morgan, J. & Pollack, M. E. (MIT Press, Cambridge, MA), pp. 417–444.
72. Hauptmann, A. G. & McAvinney, P. (1993) *Int. J. Man-Mach. Stud.* **38,** 231–249.
73. Oviatt, S. L. & Olsen, E. (1994) in *Proceedings of the International Conference on Spoken Language Processing*, eds. Shirai, K., Furui, S. & Kakehi, K. (Acoustical Society of Japan, Kobe), Vol. 2, pp. 551–554.
74. Rhyne, J. A. & Wolf, C. (1993) in *Advances in Human-Computer Interaction* (Ablex, Norwood, NJ), Vol. 4, pp. 191–250.
75. Garcia, O. N., Goldschen, A. J. & Petajan, E. D. (1992) *Technical Report* (Institute for Information Science and Technology, Department of Electrical Engineering and Computer Science, The George Washington Univ., Washington, DC).
76. Petajan, E., Bradford, B., Bodoff, D. & Brooke, N. M. (1988) in *Proceedings of Human Factors in Computing Systems (CHI'88)*, (ACM, New York), pp. 19–25.
77. Committee on Computerized Speech Recognition Technologies (1984) *Automatic Speech Recognition in Severe Environments* (Commission on Engineering and Technical Systems, National Research Council, National Academy of Sciences Press, Washington, DC).
78. Grosz, B. & Kraus, S. (1993) in *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (Chambéry, France), pp. 367–373.