Conversational Interfaces: Advances and Challenges

VICTOR W. ZUE AND JAMES R. GLASS, MEMBER, IEEE

Invited Paper

The past decade has witnessed the emergence of a new breed of human–computer interfaces that combines several human language technologies to enable humans to converse with computers using spoken dialogue for information access, creation, and processing. In this paper, we introduce the nature of these conversational interfaces and describe the underlying human language technologies on which they are based. After summarizing some of the recent progress in this area around the world, we discuss development issues faced by researchers creating these kinds of systems and present some of the ongoing and unmet research challenges in this field.

Keywords*—*Conversational interfaces, speech understanding systems, spoken dialogue systems.

I. INTRODUCTION

Computers are fast becoming a ubiquitous part of our lives, brought on by their rapid increase in performance and decrease in cost. With their increased availability comes the corresponding increase in our appetite for information. Today, for example, nearly half the population of North America are users of the World Wide Web, and the growth is continuing at an astronomical rate. Vast amounts of useful information are being made widely available, and people are utilizing it routinely for education, decision making, finance, and entertainment. Increasingly, people are interested in being able to access the information when they are on the move—anytime, anywhere, and in their native language. A promising solution to this problem, especially for small, handheld devices where a conventional keyboard and mouse can be impractical, is to impart human-like capabilities onto machines so that they can speak and hear, just like the users with whom they need to interact. Spoken language is attractive because it is the most natural, efficient, flexible, and inexpensive means of communication among humans.

When one thinks about a speech-based interface, two technologies immediately come to mind: speech recognition and speech synthesis. There is no doubt that these are important and as yet unsolved problems in their own right, with a clear set of applications that include document preparation and audio indexing. However, these technologies by themselves are often only a part of the interface solution. Many applications that lend themselves to spoken input/output—inquiring about weather or making travel arrangements—are in fact exercises in information access and/or interactive problem solving. The solution is often built up incrementally, with both the user and the computer playing active roles in the "conversation." Therefore, several language-based input and output technologies must be developed and integrated to reach this goal. The resulting *conversational interface*[1] is the subject of this paper.

Many speech-based interfaces can be considered conversational, and they may be differentiated by the degree with which the system maintains an active role in the conversation. At one extreme are *system-initiative*, or directed-dialogue, transactions, where the computer takes complete control of the interaction by requiring that the user answer a set of prescribed questions, much like the touch-tone implementation of interactive voice response (IVR) systems. In the case of air travel planning, for example, a directed-dialogue system could ask the user to "Please say just the departure city." Since the user's options are severely restricted, successful completion of such transactions is easier to attain, and indeed some successful demonstrations and deployment of such systems have been made [5].[2] At the other extreme are *user-initiative* systems, in which the user has complete freedom in what they say to the system, (e.g., "I want to visit my grandmother") while the system remains relatively passive, asking

Manuscript received January 7, 2000; revised April 25, 2000. This work was supported by DARPA under Contract N66001-99-1-8904, monitored through the Naval Command, Control and Ocean Surveillance Center.

The authors are with the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: www.sls.lcs.mit.edu).

Publisher Item Identifier S 0018-9219(00)08092-0.

[1]Throughout this paper, we will use the terms *conversational interfaces*, *conversational systems*, and *spoken dialogue systems* interchangeably.

[2]Nuance Communications: http://www.nuance.com.

C:	Yeah, [umm] I'm looking for the Buford Cinema.	*disfluency*
A:	OK, and you want to know what's showing there or ...	*interruption*
C:	Yes, please.	*confirmation*
A:	Are you looking for a particular movie?	
C:	[umm] What's showing.	*clarification*
A:	OK, one moment.	*back channel*
	...	
A:	They're showing A Troll In Central Park.	
C:	No.	*inference*
A:	Frankenstein.	*ellipsis*
C:	What time is that on?	*co-reference*
A:	Seven twenty and nine fifty.	
C:	OK, and the others?	*fragment*
A:	Little Giant.	
C:	No.	
A:	...	
C:	...	
A:	That's it.	
C:	Thank you.	
A:	Thanks for calling Movies Now.	

Fig. 1. Transcript of a conversation between an agent (A) and a client (C) over the phone. Typical conversational phenomena are annotated on the right.

Table 1 Statistics of Human–Human Dialogues in a Movie Domain [29]. Annotated Dialogue Acts are Sorted by Customer Usage and Include Frequency of Occurence and Average World Length

	Customer		Agent	
Act	Freq.	Words	Freq.	Words
Acknowledge	47.9	2.3	30.8	3.1
Request	29.5	9.0	15.0	12.3
Confirm	13.1	5.3	11.3	6.4
Inform	5.9	7.9	27.8	12.7
Statement	3.4	6.9	15.0	6.7

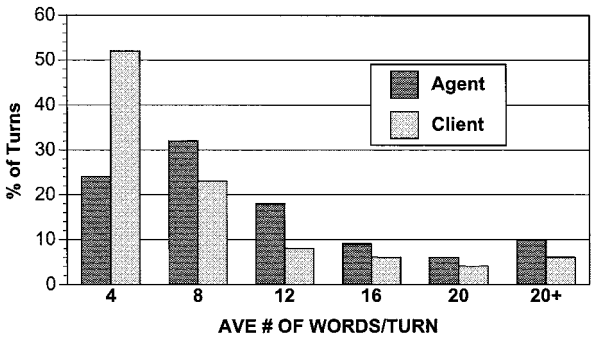

Fig. 2. Histograms of utterance length for agents and clients in tasks of information access over the phone.

only for clarification when necessary. In this case, the user may feel uncertain as to what capabilities exist, and may, as a consequence, stray quite far from the domain of competence of the system, leading to great frustration because nothing is understood. Lying between these two extremes are systems that incorporate a *mixed-initiative*, goal-oriented dialogue, in which both the user and the computer participate actively to solve a problem interactively using a conversational paradigm. It is this latter mode of interaction that is the primary focus of this paper.

What is the nature of such mixed initiative interaction? One way to answer the question is to examine human–*human* interactions during joint problem solving [29]. Fig. 1 shows the transcript of a conversation between an agent (A) and a client (C) over the phone. As illustrated by this example, spontaneous dialogue is replete with disfluencies, interruption, confirmation, clarification, ellipsis, co-reference, and sentence fragments. Some of the utterances cannot be understood properly without knowing the context in which they appear. As we shall see, while present systems cannot handle *all* these phenomena satisfactorily, some of them are being dealt with in a limited fashion.

Should one build conversational interfaces by mimicking human–human interactions? Opinion in this regard is somewhat divided. Some researchers argue that human–human dialogues can be quite variable, containing frequent interruptions, speech overlaps, incomplete or unclear sentences, incoherent segments, and topic switches. Some of these variabilities may not contribute directly to goal-directed problem solving [99]. For practical reasons, it may be desirable to ask users to modify their behavior and interact with the system in a way that is more structured. However, one may argue that users may feel more comfortable with an interface that possesses some of the characteristics of a human agent. As is the case with many other researchers, we have taken the approach of developing a human–machine interface based on analyses of human–human interactions when solving the same tasks. Regardless of the approach, we believe, as do others, that studying human–human dialogue and comparing it to human–machine dialogue can provide valuable insights [7].

Over the years, there have been many large corpora of human–human dialogues collected and analyzed (e.g., [1], [2], and [29]). For example, Table 1 shows statistics of annotated dialogue acts computed from human–human conversations in a movie information domain [29]. These statistics show that nearly half of the customers' dialogue turns were acknowledgment (e.g., "okay," "alright," "uh-huh").[3] As another example, consider the histograms of the lengths of the utterances per turn for agents and clients shown in Fig. 2 [29]. The statistics were gathered from the transcripts of more than 100 hours of conversation, in more than 1000 interactions, between agents and clients over the phone on a variety of information access tasks. More than 80% of the clients' utterances are 12 words or less, with a preponderance of very short utterances. Closer examination of the data reveals that these short utterances are mostly back-channel communications, such as "okay," "I see," etc. It is important to note that some of the spontaneous speech phenomena serve useful roles in human–human communication, and thus should conceivably be incorporated into conversational interfaces. For example, initial disfluent speech can serve an attention-getting function, and filled pauses and back-channel acknowledgment provide reassurances that the utterance is understood or one partner of the conversation is still working on the problem.

[3]An average dialogue consisted of more than 28 turns between the customer and the agent.

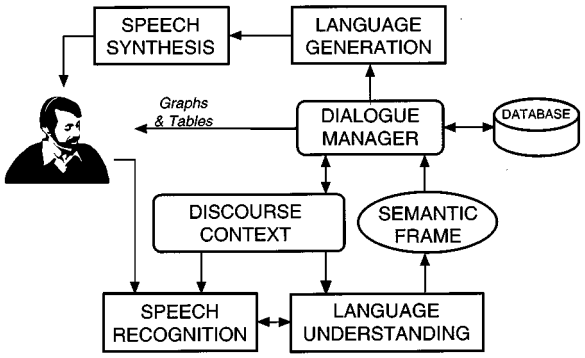

Fig. 3. Generic block diagram for a typical conversational interface.

The past decade has witnessed the emergence of some conversational systems with limited capabilities. Despite our moderate success, the ultimate deployment of such interfaces will require continuing improvement of the core human language technologies (HLTs) and the exploration into many uncharted research territories. The purpose of this paper is to outline some of these new research challenges. To set the stage, we will first introduce the components of a typical conversational system and outline some of the research issues. We will then provide a thumbnail sketch of the recent landscape, discuss some development issues concerning creation of these systems, and present some of the ongoing and unmet research challenges in this field. While we will endeavor to cover the entire field, we are unavoidably going to draw heavily from our own experience in developing such systems at MIT over the past ten years (e.g., [30], [50], [58], [87], [92], [102], [109], and [110]). This is a consequence more of familiarity then of ethnocentricity. Interested readers are referred to the recent proceedings of the Eurospeech Conference, the International Conference of Spoken Language Processing, the International Conference of Acoustics, Speech, and Signal Processing, the International Symposium on Spoken Dialogue, and other relevant publications (e.g., [22]).

II. Underlying Technologies and Research Issues

A. System Architecture

Fig. 3 shows the major components of a typical conversational interface. The spoken input is first processed through the speech-recognition component. The natural language component, working in concert with the recognizer, produces a meaning representation for the utterance. For information retrieval applications illustrated in this figure, the meaning representation can be used to retrieve the appropriate information in the form of text, tables, and graphics. If the information in the utterance is insufficient or ambiguous, the system may choose to query the user for clarification. If verbal conveyance of the information is desired, then natural language generation and text-to-speech synthesis are utilized to produce the spoken responses. Throughout the process, discourse information is maintained and fed back to the speech recognition and language understanding components, so that sentences can be properly understood in context. Finally, a dialogue component manages the interaction between the user and the computer. The nature of the dialogue can vary significantly depending on whether the system is creating or clarifying a query prior to accessing an information data base, or perhaps negotiating with the user in a post information-retrieval phase to relax or somehow modify some aspects of the initial query.

Fig. 3 does not adequately convey the notion that a conversational interface may include input and output modalities other than speech. While speech may be the interface of choice, as is the case with phone-based interactions and hands-busy/eyes-busy settings, there are clearly cases where speech is not a good modality, especially, for example, on the output side when the information contains maps, images, or large tables of information, which cannot be easily explained verbally. Human communication is inherently multimodal, employing facial, gestural, and other cues to communicate the underlying linguistic message. Thus, speech interfaces should be complemented by visual and sensory motor channels. The user should be able to choose among many modalities, including gesturing, pointing, writing, and typing on the input side [20], [88], and graphics and a talking head on the output side [55], to achieve the task in hand in the most natural and efficient manner.

The development of conversational interfaces offers a set of significant challenges to speech and natural language researchers and raises several important research issues, some of which will be discussed in the remainder of this section.

B. Spoken Input: From Signal to Meaning

Spoken language understanding involves the transformation of the speech signal into a meaning representation that can be used to interact with the specific application back-end. This is typically accomplished in two steps: the conversion of the signal to a set of words (i.e., speech recognition) and the derivation of the meaning from the word hypotheses (i.e., language understanding). A discourse component is often used to properly interpret the meaning of an utterance in the larger context of the interaction.

1) Automatic Speech Recognition: Input to conversational interfaces is often generated extemporaneously—especially from novice users of these systems. Such spontaneous speech typically contains disfluencies (i.e., unfilled and filled pauses such as “umm” and “aah,” as well as word fragments). In addition, the input utterances are likely to contain words outside the system’s working vocabulary—a consequence of the fact that present-day technology can only support the development of systems within constrained domains. Thus far, some attempts have been made to deal with the problem of disfluency. For example, researchers have improved their system’s recognition performance by introducing explicit acoustic models for the filled pauses [9], [104]. Similarly, “trash” models have been used to detect the presence of word fragments or unknown words [37], and procedures have been devised to learn the new words once they have been detected [3]. Suffice it to say, however, that the detection and learning of unknown words continues to

be a problem that needs our collective attention. A related topic is utterance- and word-level rejection, in the presence of either out-of-domain queries or unknown words [67].

An issue that is receiving increasing attention by the research community is the recognition of telephone-quality speech. It is not surprising that some of the first conversational systems available to the general public were accessible via telephone (e.g., [5], [10], and [95]), in many cases replacing presently existing IVR systems. Telephone-quality speech is significantly more difficult to recognize than high-quality recordings, because of both the limited bandwidth and the noise and distortions introduced in the channel [52]. The acoustic condition deteriorates further for cellular telephones, either analog or digital.

2) Natural Language Understanding: Speech-recognition systems typically implement linguistic constraints as a statistical language model (i.e., n-gram) that specifies the probability of a word given its predecessors. While these language models have been effective in reducing the search space and improving performance, they do not begin to address the issue of speech understanding. On the other hand, most natural language systems are developed with text input in mind; it is usually assumed that the entire word string is known with certainty. This assumption is clearly false for speech input, where many alternative words hypotheses are competing for the same time span in any sentence hypothesis produced by the recognizer (e.g., "euthanasia" and "youth in Asia,") and some words may be more reliable than others because of varying signal robustness. Furthermore, spoken language is often agrammatical, containing fragments, disfluencies, and partial words. Language understanding systems designed for text input may have to be modified in fundamental ways to accommodate spoken input.

Natural language analysis has traditionally been predominantly syntax-driven—a complete syntactic analysis is performed, which attempts to account for *all* words in an utterance. However, when working with spoken material, researchers quickly came to realize that such an approach [12], [27], [85] can break down dramatically in the presence of unknown words, novel linguistic constructs, recognition errors, and spontaneous speech events such as false starts.

Due to these problems, many researchers have tended to favor more semantic-driven approaches, at least for spoken language tasks in constrained domains. In such approaches, a meaning representation is derived by "spotting" key words and phrases in the utterance [105]. While this approach loses the constraint provided by syntax, and may not be able to adequately interpret complex linguistic constructs, the need to accommodate spontaneous speech input has outweighed these potential shortcomings. At the present time, many systems have abandoned the notion of achieving a complete *syntactic* analysis of every input sentence, favoring a more robust strategy that can still be used to produce an answer when a full parse is not achieved [42], [86], [94]. This can be accomplished by identifying parsable phrases and clauses and providing a separate mechanism for gluing them together to form a complete meaning analysis [86]. Ideally, the parser includes a probabilistic framework with a smooth transition to parsing fragments when full linguistic analysis is not achievable. Examples of systems that incorporate such *stochastic* modeling techniques can be found in [35] and [59].

How should the speech-recognition component interact with the natural language component in order to obtain the correct meaning representation? One of the most popular strategies is the so-called N-best interface [18], in which the recognizer proposes its best N complete sentence hypotheses one by one, stopping with the first sentence that is successfully analyzed by the natural language component. In this case, the natural language component acts as a filter on *whole sentence* hypotheses. Alternatively, competing recognition hypotheses can be represented in the form of a word graph [38], which is more compact than an N-best list, thus permitting a deeper search if desired.

In an N-best list, many of the candidate sentences may differ minimally in regions where the acoustic information is not very robust. While confusions such as "an" and "and" are acoustically reasonable, one of them can often be eliminated on linguistic grounds. In fact, many of the top N sentence hypotheses might be eliminated before reaching the end if syntactic and semantic analyzes take place early on in the search. One possible solution, therefore, is for the speech recognition and natural language components to be tightly coupled, so that only the acoustically promising hypotheses that are linguistically meaningful are advanced. For example, partial theories can be arranged on a stack, prioritized by score. The most promising partial theories are extended using the natural language component as a predictor of all possible next-word candidates; none of the other word hypotheses is allowed to proceed. Therefore, any theory that completes is guaranteed to parse. Researchers are beginning to find that such a tightly coupled integration strategy can achieve higher performance than an N-best interface, often with a considerably smaller stack size [32], [34], [60], [106]. The future is likely to see increasing use of linguistic analysis at earlier stages in the recognition process.

3) Discourse: Human verbal communication is a two-way process involving multiple, active participants. Mutual understanding is achieved through direct and indirect speech acts, turn taking, clarification, and pragmatic considerations. A discourse ability allows a conversational system to understand an utterance in the context of the previous interaction. As such, discourse can be considered to be part of the input processing stage. To communicate effectively, a system must be able to handle phenomena such as deictic (e.g., verbal pointing as in "I'll take the second one") and anaphoric reference (e.g., using pronouns as in "what's their phone number") to allow users to efficiently refer to items currently in focus. An effective system should also be able to handle ellipsis and fragments so that a user does not have to fully specify each query. For instance, if a user says, "I want to go from Boston to Denver," followed with, "show me only United flights," he/she clearly does not want to see *all* United flights, but rather just the ones that fly from Boston to Denver. The ability to inherit information from preceding utterances is particularly helpful in the face of recognition errors. The user may have asked a complex

question involving several restrictions, and the recognizer may have misunderstood a single word, such as a flight number or an arrival time. If a good context model exists, the user can then utter a very short correction phrase and the system will be able to replace just the misunderstood word, preventing the user from having to repeat the entire utterance, running the risk of further recognition errors.

C. Output Processing: From Information to Signal

On the output side, a conversational interface must be able to convey the information to the user in natural sounding sentences. This is typically accomplished in two steps: the information is converted into well-formed sentences, which are then fed through a text-to-speech (TTS) system to generate the verbal responses.

1) Natural Language Generation: Spoken language generation serves two important roles. First and foremost, it provides a verbal response to the user's queries, which is essential in applications where visual displays are unavailable. In addition, it can provide feedback to the user in the form of a paraphrase, confirming the system's proper understanding of the input query. Although there has been much research on natural language generation (NLG), dealing with the creation of coherent paragraphs (e.g., [56] and [75]), the language generation component of a conversational system typically produces the response one sentence at a time, without paragraph-level planning. Research in language generation for conversational systems has not received nearly as much attention as has language understanding, especially in the United States, perhaps due to the funding priorities set forth by the major government sponsors. In many cases, output sentences are simply word strings, in text or prerecorded acoustic format, that are invoked when appropriate. In some cases, sentences are generated by concatenating templates after filling slots by applying recursive rules along with appropriate constraints (person, gender, number, etc.) [31]. There has also been some recent work using more corpus-based methods for language generation in order to provide more variation in the surface realization of the utterance [63].

2) Speech Synthesis: The conversion of text to speech is the final stage of output generation. TTS systems in the past were primarily rule driven, requiring the system developers to possess extensive acoustic-phonetic and other linguistic knowledge [46]. These systems are typically very intelligible but suffer greatly in naturalness. In recent years, we have seen the emergence of a new, concatenative approach, brought on by inexpensive computation/storage and the availability of large corpora [8], [81]. In this corpus-based approach, units excised from recorded speech are concatenated to form an utterance. The selection of the units is based on a search procedure subject to a predefined distortion measure. The output of these TTS systems is often judged to be more natural than that of the rule-based systems [63].

Currently in most conversational systems, the language generation and text-to-speech components are not closely coupled; the same text is generated whether it is to be read or spoken. Furthermore, systems typically expect the language generation component to produce a textual surface form of a sentence (throwing away valuable linguistic and prosodic knowledge) and then require the text-to-speech component to produce linguistic analysis anew. Recently, there has been some work in concept-to-speech generation [57]. Such a close coupling can potentially produce higher quality output speech than could be achieved with a decoupled system, since it permits finer control of prosody. Whether language generation and speech synthesis components should be tightly integrated or can remain modular but effectively coupled by augmenting text output with a markup language (e.g., SABLE [93]) remains to be seen. Clearly, however, these two components would benefit from a shared knowledge base.

D. Dialogue Management

The dialogue modeling component of a conversational system manages the interaction between the user and the computer. The technology for building this component is one of the least developed in the HLT repertoire, especially for mixed-initiative dialogue systems considered in this paper. Although there has been some theoretical work on the structure of human–human dialogue [36], this has not led to effective insights for building human–machine interactive systems. As mentioned previously, there is also considerable debate in the speech and language research communities about whether modeling human–machine interactions after human–human dialogues is necessary or appropriate (e.g., [13], [80], and [99]).

Dialogue modeling means different things to different people. For some, it includes the *planning* and *problem solving* aspects of human–computer interactions [1]. In the context of this paper, we define dialogue modeling as the preparation, for each turn, of the system's side of the conversation, including verbal, tabular, and graphical response, as well as any clarification requests.

Dialogue modeling and management serves many roles. In the early stages of the conversation, the role of the dialogue manager might be to gather information from the user, possibly clarifying ambiguous input along the way, so that, for example, a complete query can be produced for the application data base. The dialogue manager must be able to resolve ambiguities that arise due to recognition error (e.g., "Did you say Boston or Austin") or incomplete specification (e.g., "On what day would you like to travel").

In later stages of the conversation, after information has been accessed from the data base, the dialogue manager might be involved in some negotiation with the user. For example, if there were too many items returned from the data base, the system might suggest additional constraints to help narrow down the number of choices. Pragmatically, the system must be able to initiate requests so that the information can be reduced to digestible chunks (e.g., "I found ten flights, do you have a preferred airline or connecting city").

In addition to these two fundamental operations, the dialogue manager must also inform and guide the user by suggesting subsequent subgoals (e.g., "Would you like me to price your itinerary?"), offer assistance upon request, help

relax constraints or provide plausible alternatives when the requested information is not available (e.g., "I don't have sunrise information for Oakland, but in San Francisco…"), and initiate clarification subdialogues for confirmation. In general, the overall goal of the dialogue manager is to take an active role in directing the conversation toward a successful conclusion for the user.

The dialogue manager can influence other system components by, for example, dynamically making dialogue context-dependent adjustments to language models or discourse history. At the highest level, it can help detect the appropriate broad subdomain (e.g., weather, air travel, or urban navigation). Within a particular domain, certain queries could introduce a focus of attention on a subset of the lexicon. For instance, in a dialogue about a trip to France, the initial user utterance, "I'm planning a trip to *France*," would allow the system to greatly enhance the probabilities on all the French destinations. Finally, whenever the system asks a directed question, the language model probabilities can be altered so as to favor appropriate responses to the question. For example, when the system asks the user to provide a date of travel, the system could temporarily enhance the probabilities of date expressions in the response.

There are many ways dialogue management has been implemented. Many systems use a type of scripting language as a general mechanism to describe dialogue flow (e.g., [15], [90], and [95]). Other systems represent dialogue flow by a graph of dialogue objects or modules (e.g., [5] and [98]). Another aspect of system implementation is whether or not the active vocabulary or understanding capabilities change depending on the state of the dialogue. Some systems are structured to allow a user to ask any question at any point in the dialogue so that the entire vocabulary is active at all times. Other systems restrict the vocabulary and/or language that can be accepted at particular points in the dialogue. The tradeoff is generally one of increased user flexibility (in reacting to a system response or query), and one of increased system understanding accuracy, due to the constraints on the user input.

III. Recent Progress

In the past decade there has been increasing activity in the area of conversational systems, largely due to government funding in the United States and Europe. By the late 1980s, the DARPA spoken language systems (SLS) program was initiated in the U.S., while the Esprit SUNDIAL (Speech Understanding and dialogue) program was under way in Europe [69]. The task domains for these two programs were remarkably similar in that both involved data base access for travel planning, with the European one including both flight and train schedules and the American one being restricted to air travel. The European program was a multilingual effort involving four languages (English, French, German, and Italian), whereas the American effort was, understandably, restricted to English. All of the systems focused within a narrowly defined area of expertise, and vocabulary sizes were generally limited to several thousand words. Nowadays, these types of systems can typically run in real-time on standard workstations and PCs with no additional hardware.

Strictly speaking, the DARPA SLS program cannot be considered conversational in that its attention focused entirely on the input side. However, since the technology developed during the program had a significant impact on the speech understanding methods used by conversational systems, it is worth describing in more detail. The program adopted the approach of developing the underlying input technologies within a common domain called Air Travel Information Service (ATIS) [74]. ATIS permits users to query for air travel information, such as flight schedules from one city to another, obtained from a small, static relational data base excised from the Official Airline Guide. By requiring that all system developers use the same data base, it was possible to compare the performance of various spoken language systems based on their ability to extract the correct information from the data base, using a set of prescribed training and test data and a set of interpretation guidelines. Indeed, common evaluations occurred at regular intervals, and steady performance improvements were observed for all systems. At the end of the program, the best system achieved a word error rate of 2.3% and a sentence error rate of 15.2% [66]. Additionally, the best system achieved an understanding error rate of 5.9% and 8.9% for text and speech input, respectively.[4]

The European SUNDIAL project differed in several ways from the DARPA SLS program. Whereas the SLS program had regular common evaluations, the SUNDIAL project had none. Unlike the SLS program however, the SUNDIAL project aimed at building systems that could be publicly deployed. For this reason, the SUNDIAL project designated dialogue modeling and spoken language generation as integral parts of the research program. As a result, this has led to some interesting advances in Europe in dialogue control mechanisms.

Since the end of the SLS and SUNDIAL programs in 1995 and 1993, respectively, there have been other sponsored programs in spoken dialogue systems. In the recently completed Automatic Railway Information Systems for Europe (ARISE) project, which was a part of the LE3 program, participants developed train timetable information systems covering three different languages (Dutch, French, and Italian) [23]. Groups explored alternative dialogue strategies and investigated different technology issues. Four prototypes underwent substantial testing and evaluation (e.g., [17], [49], and [82]). In the United States, a new DARPA funded project called Communicator has begun, which emphasizes dialogue-based interactions incorporating both speech input and output technologies. One of the properties of this program is that participants are using a common system architecture to encourage component sharing across sites [89]. Participants in this program are developing both their own dialogue domains and a common complex travel task (e.g., [28]).

[4]All the performance results quoted here are for the "evaluable" queries, i.e., those queries that are within domain and for which an appropriate answer is available from the data base.

Table 2 A Comparison of Several Conversational Systems that have been Deployed and Used by Real Users

Domain	Language	Vocabulary Size	Average Words/Utt	Average Utts/Dialogue
CSELT Train Timetable Info	Italian	760	1.6	6.6
SpeechWorks Air Travel Reservation	English	1000	1.9	10.6
Philips Train Timetable Info	German	1850	2.7	7.0
CMU Movie Information	English	757	3.5	9.2
CMU Air Travel Reservation	English	2851	3.6	12.0
LIMSI Train Timetable Info	French	1800	4.4	14.6
MIT Weather Information	English	1963	5.2	5.6
MIT Air Travel Reservation	English	1100	5.3	14.1
AT&T Operator Assistance	English	4000	7.0	3.0
Air Travel Reservations (human)	English	?	8.0	27.5

In addition to the research sponsored by these larger programs, there have been many other independent initiatives as well. Although there are far too many to list here, some examples include the Berkeley Restaurant Project (BeRP), which provided restaurant information in the Berkeley, CA, area [43]. The AT&T AutoRes system allowed users to make rental car reservations over the phone via a toll-free number [54]. Their "How may I help you?" system provides call routing services and information [35]. The Waxholm system provides ferry timetables and tourist information for the Stockholm archipelago [11]. At the University of Rochester, the TRAINS project involved train schedule planning [1].

One of the most noticeable trends in spoken dialogue systems is the increasing number of publicly deployed systems. Such systems include not only research prototypes but also commercial products which are used on a much wider scale for domains such as call routing, stock quotes, train schedules, and flight reservations. (e.g., [4], [5], and [10]).

Although it can be difficult to compare different systems, it is interesting to observe some of their basic properties. Table 2 shows some statistics of several different systems which have been deployed and used by real users. Each system is characterized in terms of the domain of operation, language, vocabulary size, and the average number of words per utterance and utterances per dialogue. The systems are listed in increasing order of average number of words per utterance. The first three systems are examples of commercial products and/or have been deployed on a very large scale (i.e., fielding millions of calls): the CSELT train timetable information system [10], the SpeechWorks air travel reservation system [5], and the Philips TABA train timetable information system [95]. The second group of six systems are examples of research prototypes which have been made publicly available on a smaller scale. They include movie information and air travel reservation systems developed at CMU[5] [79], the LIMSI train timetable information system [76], weather and air travel information systems developed at MIT [91], [110], and the AT&T "How may I help you?" operator assistance system [35]. The final statistics were computed from a set of 66 air travel reservation transactions between customers and agents, which were transcribed by SRI [47].

[5]CMU Movieline: http://www.speech.cs.cmu.edu/Movieline.

From the table, we can see that most of these systems have vocabulary sizes in the thousands, although for other domains such as stock quotes, the vocabulary size could be considerably larger. It is interesting to observe that the average number of words per utterance tends to increase as one moves from commercial systems, to research prototypes, to human–human dialogues. Naturally, there are many factors that affect averages, including the basic nature of the application. However, it is likely that systems that employ more system-initiative or directed dialogues (by asking the user to answer specific questions) or that require explicit confirmations would also tend to have fewer words per utterance on average. It is also apparent that none of the human–machine dialogues was as wordy as those between humans.

IV. Development Issues

Spoken dialogue systems require first and foremost the availability of high-performance human language technology components such as speech recognition and language understanding. However, the development of these systems also demands that we pay close attention to a host of other issues. While many of these issues may have little to do with human language technologies *per se*, they are nonetheless crucial to successful system development. In this section, we will outline some of these development issues.

A. Working in Real Domains

The objective for developing a conversational interface is to provide a natural way for any user, especially the computer illiterate, to access and manage information. Since humans will ultimately be consumers of this technology, it is important that the systems be developed with their behaviors and needs in mind. An effective strategy, and one that we subscribe to, involves the development of the underlying technologies within *real* application domains, rather than relying on artificial scenarios, however realistic they might be. Such a strategy will force us to confront some of the critical research issues that may otherwise elude our attention, such as dialogue modeling, new word detection/learning, confidence scoring, robust recognition of accented speech, and portability across domains and languages. We also believe that working on real applications has the potential benefit of shortening the interval between technology demonstration and its deployment. Above all, real applications that can help

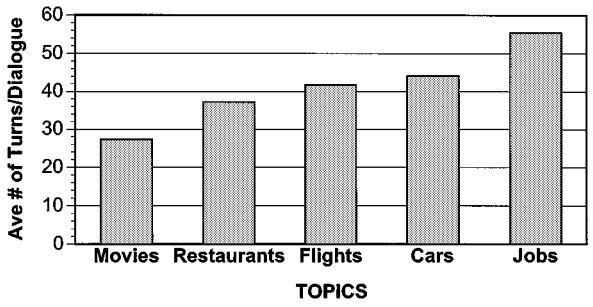

Fig. 4. Averaged number of dialogue turns for several application domains.

people solve problems will be used by real users, thus providing us with a rich and continuing source of useful data. These data are far more useful than anything we could collect in a laboratory environment.

What constitutes real domains and real users? One may rightfully argue that only commercially deployed systems capable of providing robust and scalable solutions can truly be considered real. A laboratory prototype that is only available to the public via perhaps a single phone line is quite different from a commercially deployed system. Nevertheless, we believe a research strategy that incorporates as much realism as possible early into the system's research and development life cycle is far more preferable to one that attempts to develop the underlying technologies in a concocted scenario. At the very least, a research prototype capable of providing real and useful information, made available to a wide range of users, offers a valuable mechanism for collecting data that will benefit the development of both types of systems. We will illustrate this point in the next section with one of the systems we have developed at MIT.

How do we select the applications that are well matched to our present capabilities? The answer may lie in examining human–human data. Fig. 4 displays the average number of dialogue turns per transaction for several application domains. The data are obtained from the same transcription of the 100 hours of real human–human interactions described earlier. As the data clearly show, helping a user select a movie or a restaurant is considerably less complex than helping a user look for employment.

B. Data Collection

Developing conversational interfaces is a classic chicken and egg problem. In order to develop the system capabilities, one needs to have a large corpus of data for system development, training, and evaluation. In order to collect data that reflect actual usage, one needs to have a system that users can speak to. Fig. 5 illustrates a typical cycle of system development. For a new domain or language, one must first develop some limited natural language capabilities, thus enabling an "experimenter-in-the-loop," or *wizard-of-oz*, data collection paradigm, in which an experimenter types the spoken sentences to the system after removing spontaneous speech artifacts. This process has the advantage of eliminating potential recognition errors. The resulting data are then used for the development and training of the speech recognition and

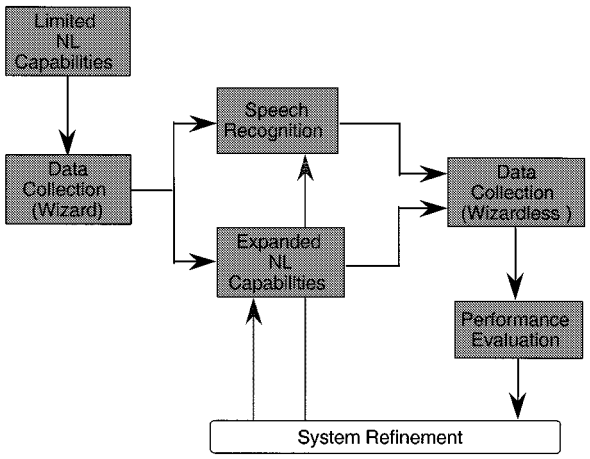

Fig. 5. Illustration of data collection procedures.

natural language components. As these components begin to mature, it becomes feasible to collect more data using the "system-in-the-loop," or *wizardless*, paradigm, which is both more realistic and more cost effective. Performance evaluation using newly collected data will facilitate system refinement.

The means and scale of data collection for system development and evaluation have evolved considerably over the past decade. This is true for both the speech recognition and speech understanding communities, and can be seen in many of the systems in the recent ARISE project [23] and elsewhere. At MIT, for example, the Voyager urban navigation system was developed in 1989 by recruiting 100 subjects to come to our laboratory and ask a series of questions to an initial wizard-based system [30]. In contrast, the data collection procedure for the more recent Jupiter weather information system consists of deploying a publicly available system and recording the interactions [110]. There are large differences in the number of queries, the number of users, and the range of issues that the data provide. By using a system-in-the-loop form of data collection, system development and evaluation become iterative procedures. If unsupervised methods were used to augment the system ASR and NLU capabilities, system development could become continuous (e.g., [45]).

Fig. 6 shows, over a two-year period, the cumulative amount of data collected from real users using the MIT Jupiter system and the corresponding word error rates (WERs) of our recognizer. Before we made the system accessible through a toll-free number, the WER was about 10% for laboratory collected data. The WER more than tripled during the first week of data collection. As more data were collected, we were able to build better lexical, language, and acoustic models. As a result, the WER continued to decrease over time. This negative correlation suggests that making the system available to real users is a crucial aspect of system development. If the system can provide real and useful information to users, they will continue to call, thus providing us with a constant supply of useful data. However, in order to get users to actually use the system, it needs to be providing "real" information to the user. Otherwise, there

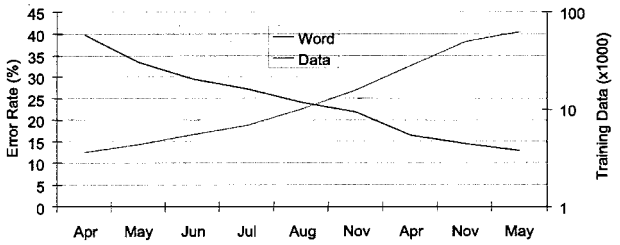

Fig. 6. Comparison of recognition performance and the number of utterances collected from real users over time in the MIT weather domain. Note that the x-axis has a nonlinear time scale, reflecting the time when new versions of the recognizer were released.

is little incentive for people to use the system other than to play around with it, or to solve toy problem scenarios that may or may not reflect problems of interest to real users.

C. Evaluation

One of the issues that developers of spoken dialogue systems face is how to evaluate progress, in order to determine if they have created a usable system. Developers must decide what metrics to use to evaluate their systems to ensure that progress is being made. Metrics can include component evaluations but should also assess the overall performance of their system.

For systems that conduct a transaction, it is possible to tell whether or not a user has completed a task. In these cases, it is also possible to measure accompanying statistics such as the length of time to complete the task, the number of turns, etc. It has been noted, however, that such statistics may not be as important as user satisfaction (e.g., [77]). For example, a spoken dialogue interface may take longer than some alternative, yet users may prefer it due to other factors (less stressful, hands-free, etc). A better form of evaluation might be a measure of whether users liked the system, whether they called to perform a real task (rather than browsing), and whether they would use it again, or recommend it to others. Evaluation frameworks such as Paradise [100] attempt to correlate system measurements with user satisfaction in order to better quantify these effects [101].

Although there have been some recent efforts in evaluating language output technologies (e.g., TTS comparisons [83]), evaluation methods for ASR and NLU have been more common since they are more amenable to automatic evaluation methods where it is possible to decide what is a correct answer. ASR evaluation has tended to be the most straightforward, although there are a range of phenomena that are not necessarily obvious how to evaluate (e.g., crosstalk, mumbling, partial words). NLU evaluation can also be performed by comparing some form of meaning representation with a reference. In [73], for example, two metrics are measured on an utterance-by-utterance basis, which attempt to assess the performance of discourse and dialogue in addition to ASR and NLU. The first measures the average number of attributes introduced per query (a measure of information rate), while the second measures how many turns it took, on average, for an intended attribute to be transmitted successfully to the system (a measure of user frustration).

One problem with NLU evaluation is that there is no common meaning representation among different research sites, so cross-site comparison becomes difficult. In the DARPA SLS program, for example, the participants ultimately could agree only on comparing to an answer coming from a common data base. Unfortunately, this necessarily led to the creation of a large document defining principals of interpretation for all conceivable queries [40]. In order to keep the response across systems consistent, systems were restricted from taking the initiative, which is a major constraint on dialogue research.

One way to show progress for a particular system is to perform longitudinal evaluations for recognition and understanding. In the case of Jupiter, as shown in Fig. 6, we continually evaluate on standard test sets, which we can redefine periodically in order to keep from tuning to a particular data set [72], [110]. Since data continually arrive, it is not difficult to create new sets and reevaluate older system releases on these new data.

Some systems make use of dialogue context to provide constraints for recognition, e.g., favoring candidate hypotheses that mention a date after the system has just asked for a date. Thus, any reprocessing of utterances in order to assess improvements in recognition or understanding performance at a later time need to be able to take advantage of the *same* dialogue context as was present in the original dialogue with the user. To do this, the dialogue context must be recorded at the time of data collection and reutilized in the subsequent off-line processing, in order to avoid giving the original system an unwarranted advantage [73].

V. Challenges

As we can see, considerable progress has been made over the past decade in research and development of systems that can understand and respond to spoken language. To meet the challenges of developing a language-based interface to help users solve real problems, however, we must continue to improve the core technologies while expanding the scope of the underlying human language technology base. In this section, we highlight some of the new research challenges that deserve our collective attention, realizing that the list is but a sampling of the entire landscape.

A. Spoken Language Understanding

The development of conversational systems shares many of the research challenges being addressed by the speech recognition community for other applications such as speech dictation and spoken document retrieval, although the recognizer is often exercised in different ways. For example, in contrast to desktop dictation systems, the speech recognition component in a conversational system is often required to handle a wide range of channel variations. Increasingly, landline and cellular phones are the transducer of choice, thus requiring the system to deal with narrow channel bandwidths, low signal-to-noise ratios, diversity in handset characteristics, dropout, and other artifacts. In many situations where speech input is especially appropriate (e.g., hands-busy/eyes-

busy), there can be significant background noise (e.g., cars) and possibly stress on the part of the speaker. Robust conversational systems will be required to handle these types of phenomena.

Another problem that is particularly acute for conversational systems is the recognition of speech from a diverse speaker population. In the data we collected for Jupiter, for example, we observed a significant number of children, as well as users, with strong dialects and nonnative accents. The challenge posed by these data to speaker-independent recognition technology must be met [53], since conversational interfaces are intended to serve people from all walks of life.

A solution to these channel and speaker variability problems may be adaptation. For applications in which the entire interaction consists of only a few queries, short-term adaptation using only a small amount of data would be necessary. For applications where the user identity is known, the system can make use of user profiles to adapt not only acoustic-phonetic characteristics but also pronunciation, vocabulary, language, and possibly domain preferences (e.g., user lives in Boston, prefers aisle seat when flying).

An important problem for conversational systems is the detection and learning of new words. In a domain such as Jupiter or electronic Yellow Pages, a significant fraction of the words uttered by users may not be in the system's working vocabulary. This is unavoidable partly because it is not possible to anticipate all the words that all users are likely to use, and partly because the data base is usually changing with time (e.g., new restaurants opening up). In systems such as Jupiter, users will sometimes try to help the system with unknown city names by spelling the word (e.g., "I said B A N G O R, Bangor"), or emphasizing the syllables in the word (which usually leads to worse results). In the past, we have not paid much attention to the unknown word problem because the tasks the speech-recognition community has chosen often assume a closed vocabulary. In the limited cases where the vocabulary has been open, unknown words have accounted for a small fraction of the word tokens in the test corpus. Thus, researchers could either construct generic "trash word" models and hope for the best, or ignore the unknown word problem altogether and accept a small penalty on word error rate. In real applications, however, the system must be able to cope with unknown words simply because they will always be present, and ignoring them will not satisfy the user's needs—if a person wants to know how to go from the train station to a restaurant whose name is unknown to the system, they will not settle for a response such as, "I am sorry, I don't understand you. Please rephrase the question." The system must be able not only to *detect* new words, taking into account acoustic, phonological, and linguistic evidence, but also to adaptively *acquire* them, in terms of both their orthography and linguistic properties. In some cases, fundamental changes in the problem formulation and search strategy may be necessary. While some research is being conducted in this area [3], [37], much more work remains to be done.

For simple applications such as auto-attendant, it is possible for a conversational system to achieve "understanding" without utilizing sophisticated natural language processing techniques. For example, one could perform keyword or phrase spotting on the recognizer's output to obtain a meaning representation. As the interactions become more complex, involving multiple turns, the system may need more advanced natural language analysis in order to achieve understanding in context.

Although there are many examples in the literature of both partial and fully unsupervised learning methods applied to natural language processing, NLU in the context of conversational systems remains mainly a knowledge intensive process. Even stochastic approaches that can learn the linguistic regularities automatically require that a large corpus be properly annotated with syntactic and semantic tags [35], [59], [68]. One of the continuing challenges facing researchers is the discovery of processes that can automate the discovery of linguistic facts.

Competing strategies to achieve robust understanding have been explored in the research community. For example, the system could adopt the strategy of first performing word- and phrase-spotting and rely on full linguistic analysis only when necessary. Alternatively, the system could first perform full linguistic analysis in order to uncover the linguistic structure of the utterance and relax the constraints through robust parsing and word/phrase-spotting only when full linguistic analysis fails. At this point, it is not clear which of these strategies would yield the best performance. Continued investigation is clearly necessary.

B. Spoken Language Generation

With few exceptions, current research in spoken language systems has focused on the input side, i.e., the understanding of the input queries, rather than the *conveyance* of the information. It is interesting to observe, however, that the speech synthesis component is the one that often leaves the most lasting impression on users—especially when it does not sound especially natural. As such, more natural sounding speech synthesis will be an important research topic for spoken dialogue systems in the future.

Spoken language generation is an extremely important aspect of the human–computer interface problem, especially if the transactions are to be conducted over a telephone. Models and methods must be developed that will generate natural sentences appropriate for spoken output across many domains and languages. For applications where all information must be conveyed aurally, particular attention must be paid to the interaction between language generation and dialogue management—the system may have to initiate a clarification subdialogue to reduce the amount of information returned from the back-end, in order not to generate unwieldy verbal responses.

As mentioned earlier, recent work in speech synthesis based on nonuniform units has resulted in much improved synthetic speech quality [41], [81]. However, we must continue to improve speech synthesis capabilities, particularly with regard to the encoding of prosodic and possibly paralinguistic information such as emotion. As is the case

on the input side, we must also explore integration strategies for language generation and speech synthesis. Finally, evaluation methodologies for spoken language generation technology must be continued to be developed and more comparative evaluations performed [83].

Many researchers have observed that the precise wording of the system response can have a large impact on the user response. In general, the more vaguely worded response will result in the larger variation of inputs [5], [76]. Which type of response is more desirable will perhaps depend on whether the system is used for research or commercial purposes. If the final objective is to improve understanding of a wider variety of input, then a more general response might be more appropriate. A more directed response, however, would most likely improve performance in the short term.

The language generation used by most spoken dialogue systems tends to be static, using the identical response pattern in its interaction with users. While it is quite possible that users will prefer consistent feedback from a system, we have observed that introducing variation in the way we prompt users for additional queries (e.g., "Is there anything else you'd like to know?" "Can I help you with anything else?" "What else?") is quite effective in making the system appear less robotic and more natural to users. It would be interesting to see if a more stochastic language generation capability would be well received by users. In addition, the ability to vary the prosody of the output (e.g., apply contrastive stress to certain words) also becomes important in reducing the monotony and unnaturalness of speech responses.

A more philosophical question for language generation is whether or not to personify the system in its responses to users. Naturally, there are varied opinions on this matter. In many situations we have found that an effective response is one commonly used in human–human interaction (e.g., "I'm sorry"). Users do not seem to be bothered by the personification evident in our deployed systems.

Although prosody impacts both speech understanding and speech generation, prosodic features have been most widely incorporated into text-to-speech systems. However, there have been attempts to make use of prosodic information for both recognition and understanding [39], [64], [84], and it is hopeful that more research will appear in this area in the future. In the Verbmobil project, researchers have been able to show considerable improvement in processing speed when integrating prosodic information into the search component during recognition [62].

C. Dialogue Management

In most current dialogue systems, the design of the dialogue strategy is typically hand-crafted by the system developers, and as such is largely based on their intuition about the proper dialogue flow. This can be a time-consuming process, especially for mixed-initiative dialogues, whose result may not generalize to different domains. There has been some recent research exploring the use of machine learning techniques to automatically determine dialogue strategy [51]. Regardless of the approach, however, there is the need to develop the necessary infrastructure for dialogue research. This includes the collection of dialogue data, both human–human and human–machine. These data will need to be annotated, after developing annotation tools and establishing proper annotation conventions. In the past decade, speech recognition and language understanding communities have benefited from the availability of large, annotated corpora. Similar efforts are desperately needed for dialogue modeling. Organizations such as the Special Interest Group on Dialogue (SIGdial) of the Association for Computational Linguistics aim to advance dialogue research in areas such as portability, evaluation, resource sharing, and standards, among others.[6]

Since we are far from being able to develop omnipotent systems capable of unrestricted dialogue, it is necessary for current systems to accurately convey their limited capabilities to the user, including both the domain of knowledge of the system itself and the kind of speech queries that the system can understand. While expert users can eventually become familiar with at least a subset of the system capabilities, novices can have considerable difficulty if their expectations are not well matched with the system capabilities. This issue is particularly relevant for mixed-initiative dialogue systems; by providing more flexibility and freedom to users to interact with the system, one could potentially increase the danger of them straying out of the system's domain of expertise. For example, our Jupiter system knows only short-term weather forecasts, yet users ask a wide variety of legitimate weather questions (e.g., "What's the average rainfall in Guatemala in January?" or, "When is high tide tomorrow?") that are outside the system's capabilities, along with a wide variety of nonweather queries. Even if users are aware of the system's domain of knowledge, they may not know the *range* of knowledge within the domain. For example, Jupiter does not know all 23 000 cities in the United States, so it is necessary to be able to detect when a user is asking for an out-of-vocabulary city, and then help inform the user what cities the system knows without listing all possibilities. Finally, even if the user knows the full range of capabilities of the system, he/she may not know what type of questions the system is able to understand.

In order to assist users to stay within the capabilities of the system, some form of "help" capability is required. However, it is difficult to provide help capabilities since users may not know when to ask for it, and when they do, the help request may not be explicit, especially if they do not understand why the system was misbehaving in the first place. Regardless, the help messages will clearly need to be context dependent, with the system offering the appropriate suggestions depending on the dialogue states.

Another challenging area of research is the recovery from the inevitable misunderstandings that a system will make. Errors could be due to many different phenomena (e.g., acoustics, speaking style, disfluencies, out-of-vocabulary words, parse coverage, or understanding gaps), and it can be difficult

[6]SIGdial: http://www.sigdial.org/.

to detect that there is a problem, determine what the problem is caused by, and convey to the user an appropriate response that will fix the problem.

Many systems incorporate some form of confidence scoring to try to identify problematic inputs (e.g., [5] and [44]). The system can then either try an alternative strategy to help the user, or back off to a more directed dialogue and/or one that requires explicit confirmation [76], [79], [96]. Based on our statistics with Jupiter, however, we have found that when an utterance is rejected, it is highly likely that the next utterance will be rejected as well [67]. Thus, it appears that certain users have an unfortunate tendency to go into a rejection death spiral that can be hard to get out of. Using confidence scoring to perform partial understanding might allow for more refined corrective dialogue (e.g., requesting input of only the uncertain regions). Partial understanding may also help in identifying out-of-vocabulary words and enable more constructive feedback from the system about the possible courses of action (e.g., "I heard you ask for the weather in a city in New Jersey. Can you spell it for me?").

Spoken dialogue systems can behave quite differently depending on what input and output modalities are available to the user. In displayless environments such as the telephone, it might be necessary to tailor the dialogue so as not to overwhelm the user with information. When displays are available, however, it may be more desirable to simply summarize the information to the user, and to show them a table or image, etc. Similarly, the nature of the interaction will change if alternative input modalities, such as pen or gesture, are available to the user. Which modality is most effective will depend, among other things, on environment (e.g., classroom), user preference, and perhaps dialogue state [65].

Researchers are also beginning to study the addition of back-channel communication in spoken dialogue responses in order to make the interaction more natural. Prosodic information from fundamental frequency and duration appear to provide important clues as to when back-channeling might occur [61], [107]. Intermediate feedback from the system can also be more informative to the user than silence or idle music when inevitable delays occur in the dialogue (e.g., "Hold on while I look for the cheapest price for your flight to London…").

Finally, many systems are able to handle interruptions by allowing the user to "barge in" over the system response (e.g., [5], [79]). To date, barge-in has been treated primarily as an acoustic problem, with perhaps some interaction with a speech recognizer. However, it clearly should also be viewed as an understanding problem, so that the system can differentiate among different types of input such as noise, back-channel, or a significant question or statement, and take appropriate actions. In addition, it will be necessary to properly update the dialogue status to reflect the fact that barge-in occurred. For example, if the system was reading a list of flights, the system might need to remember where the interruption occurred—especially if the interruption was underspecified (e.g., "I'll take the United flight" or "Tell me about that one").

D. Portability

Creating a robust, mixed-initiative dialogue system can require a tremendous amount of effort on the part of researchers. In order for this technology to ultimately be successful, the process of porting existing technology to new domains and languages must be made easier. Over time, researchers have made the technology more modular.

Over the past few years, different research groups have been attempting to make it easier for nonexperts to create new domains. Systems that modularize their dialogue manager try to take advantage of the fact that a dialogue can often be broken down into a set of smaller subdialogues (e.g., dates, addresses), in order to make it easier to construct dialogue for a new domain (e.g., [5] and [98]). For example, researchers at OGI have developed rapid development kits for creating spoken dialogue systems, which are freely available [98] and which have been used by students to create their own systems [97]. On the commercial side, there has been a significant effort to develop the Voice eXtensible Markup Language (VoiceXML) as a standard to enable internet content and information access via voice and phone.[7] To date, these approaches have been applied only to directed dialogue strategies. Much more research is needed in this area if we are to try to allow systems with complex dialogue strategies to generalize to different domains.

Currently, the development of speech recognition and language understanding technologies has been domain and language specific, requiring a large amount of annotated training data. However, it may be costly, or even impossible, to collect a large amount of training data for certain applications or languages. Therefore, we must address the problems of producing a conversational system in a new domain and language given at most a small amount of domain-specific training data. To achieve this goal, we must strive to cleanly separate the algorithmic aspects of the system from the application-specific aspects. We must also develop automatic or semiautomatic methods for acquiring the acoustic models, language models, grammars, semantic structures for language understanding, and dialogue models required by a new application. The issue of portability spans across different acoustic environments, data bases, knowledge domains, and languages. Real deployment of multilingual spoken language technology cannot take place without adequately addressing this issue.

VI. Concluding Remarks

In this paper, we have attempted to outline some of the important research challenges that must be addressed before spoken language technologies can be put to pervasive use. The timing for the development of human language technology is particularly opportune, since the world is mobilizing to develop the information highway that will be the backbone of future economic growth. Human language technology will play a central role in providing an interface that

[7]VoiceXML: http://www.voicexml.org/.

will dramatically change the human–machine communication paradigm from *programming* to *conversation*. It will enable users to efficiently access, process, manipulate, and absorb a vast amount of information. While much work needs to be done, the progress made collectively by the community thus far gives us every reason to be optimistic about fielding such systems, albeit with limited capabilities, in the near future.

ACKNOWLEDGMENT

The authors would like to thank P. Baggia, A. Halberstadt, L. Lamel, G. Riccardi, A. Rudnicky, and B. Souvignier for providing the statistics about their spoken dialogue systems used in this paper. They would also like to thank J. Polifroni, S. Seneff, and the two anonymous reviewers who made many helpful suggestions to improve the overall quality of this paper.

REFERENCES

[1] J. Allen *et al.*, "The TRAINS project: A case study in defining a conversational planning agent," *J. Exper. Theoret. AI*, vol. 7, pp. 7–48, 1995.

[2] A. Anderson *et al.*, "The HCRC map task corpus," *Lang. Speech*, vol. 34, no. 4, pp. 351–366, 1992.

[3] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large vocabulary continuous speech recognition system," in *Proc. ICASSP*, 1991, pp. 305–308.

[4] H. Aust, M. Oerder, F. Seide, and V. Steinbiss, "The Philips automatic train timetable information system," *Speech Commun.*, vol. 17, pp. 249–262, 1995.

[5] E. Barnard, A. Halberstadt, C. Kotelly, and M. Phillips, "A consistent approach to designing spoken-dialog systems," presented at the Proc. ASRU Workshop, Keystone, CO, 1999.

[6] M. Bates, R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwartz, and D. Stallard, "The BBN/HARC spoken language understanding system," in *Proc. ICASSP*, 1993, pp. 111–114.

[7] N. Bernsen, L. Dybkjaer, and H. Dybkjaer, "Cooperativity in human–machine and human–human spoken dialogue," *Discourse Processes*, vol. 21, no. 2, pp. 213–236, 1996.

[8] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," presented at the Proc. ASA, Berlin, Germany, 1999.

[9] J. Butzberger, H. Murveit, and M. Weintraub, "Spontaneous speech effects in large vocabulary speech recognition applications," in *Proc. ARPA Workshop Speech and Natural Language*, 1992, pp. 339–344.

[10] R. Billi, R. Canavesio, and C. Rullent, "Automation of telecom italia directory assistance service: Field trial results," in *Proc. IVTTA*, 1998, pp. 11–16.

[11] M. Blomberg, R. Carlson, K. Elenius, B. Granstrom, J. Gustafson, S. Hunnicutt, R. Lindell, and L. Neovius, "An experimental dialogue system: Waxholm," in *Proc. Eurospeech*, 1993, pp. 1867–1870.

[12] R. Bobrow, R. Ingria, and D. Stallard, "Syntactic and semantic knowledge in the DELPHI unification grammar," in *Proc. DARPA Speech Natural Language Workshop*, 1990, pp. 230–236.

[13] L. Boves and E. den Os, "Applications of speech technology: Designing for usability," in *Proc. IEEE Workshop ASR and Understanding*, 1999, pp. 353–362.

[14] B. Buntschuh *et al.*, "VPQ: A spoken language interface to large scale directory information," in *Proc. ICSLP*, 1998, pp. 2863–2866.

[15] R. Carlson and S. Hunnicutt, "Generic and domain- specific aspects of the waxholm NLP and dialogue modules," in *Proc. ICSLP*, 1996, pp. 677–680.

[16] J. Cassell, "Embodied conversation: Integrating face and gesture into automatic spoken dialogue systems," in Spoken Dialogue Systems, Luperfoy, Ed. Cambridge, MA: MIT Press, to be published.

[17] G. Castagnieri, P. Baggia, and M. Danieli, "Field trials of the Italian ARISE train timetable system," in *Proc. IVTTA*, 1998, pp. 97–102.

[18] Y. Chow and R. Schwartz, "The N-best algorithm: An efficient procedure for finding top N sentence hypotheses," in *Proc. ARPA Workshop Speech and Natural Language*, 1989, pp. 199–202.

[19] M. Cohen, Z. Rivlin, and H. Bratt, "Speech recognition in the ATIS domain using multiple knowledge sources," in *Proc. DARPA Spoken Language Systems Technology Workshop*, 1995, pp. 257–260.

[20] P. Cohen, M. Johnson, D. McGee, S. Oviatt, J. Clow, and I. Smith, "The efficiency of multimodal interaction: A case study," in *Proc. ICSLP*, 1998, pp. 249–252.

[21] P. Constantinides, S. Hansma, and A. Rudnicky, "A Schema-based approach to dialog control," in *Proc. ICSLP*, 1998, pp. 409–412.

[22] P. Dalsgaard, L. Larsen, and I. Thomsen, Eds., *Proc. ESCA Tutorial and Research Workshop Spoken Dialogue Systems: Theory and Application* Vigsø, Denmark, 1995.

[23] E. den Os, L. Boves, L. Lamel, and P. Baggia, "Overview of the ARISE project," in *Proc. Eurospeech*, 1999, pp. 1527–1530.

[24] M. Denecke and A. Waibel, "Dialogue strategies guiding users to their communicative goals," in *Proc. Eurospeech*, 1997, pp. 2227–2230.

[25] L. Devillers and H. Bonneau-Maynard, "Evaluation of dialog strategies for a tourist information retrieval system," in *Proc. ICSLP*, 1998, pp. 1187–1190.

[26] V. Digilakis *et al.*, "Rapid speech recognizer adaptation to new speakers," in *Proc. ICASSP*, 1999, pp. 765–768.

[27] J. Dowding, J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken language understanding," in *Proc. ARPA Workshop on Human Language Technology*, 1993, pp. 21–24.

[28] M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, C. Bennett, and J. Allen, "Data collection and processing in the Carnegie Mellon communicator," in *Proc. Eurospeech*, 1999, pp. 2695–2698.

[29] G. Flammia, "Discourse segmentation of spoken dialogue: An empirical approach," Ph.D. dissertation, MIT, Cambridge, MA, 1998.

[30] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT voyager system," *Speech Commun.*, vol. 17, pp. 1–18, 1995.

[31] J. Glass, J. Polifroni, and S. Seneff, "Multilingual language generation across multiple domains," in *Proc. ICSLP*, 1994, pp. 983–976.

[32] D. Goddeau, "Using probabilistic shift-reduce parsing in speech recognition systems," in *Proc. ICSLP*, 1992, pp. 321–324.

[33] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai, "A form-based dialogue manager for spoken language applications," in *Proc. ICSLP*, 1996, pp. 701–704.

[34] D. Goodine, S. Seneff, L. Hirschman, and M. Phillips, "Full integration of speech and language understanding in the MIT spoken language system," in *Proc. Eurospeech*, 1991, pp. 845–848.

[35] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.

[36] B. Grosz and C. Sidner, "Plans for discourse," in *Intentions in Communication*. Cambridge, MA: MIT Press, 1990.

[37] L. Hetherington and V. Zue, "New words: Implications for continuous speech recognition," in *Proc. Eurospeech*, 1991, pp. 475–931.

[38] L. Hetherington, M. Phillips, J. Glass, and V. Zue, "A^* word network search for continuous speech recognition," in *Proc. Eurospeech*, 1993, pp. 1533–1536.

[39] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," in *Proc. ESCA Workshop Dialogue and Prosody*, 1999, pp. 7–15.

[40] L. Hirschman *et al.*, "Multi-site data collection for a spoken language corpus," in *Proc. DARPA Workshop Speech and Natural Language*, 1992, pp. 7–14.

[41] X. Huang, A. Acero, J. Adcock, H. W. Hon, J. Goldsmith, J. Liu, and M. Plumpe, "WHISTLER: A trainable text-to-speech system," in *Proc. ICSLP*, 1996, pp. 2387–2390.

[42] E. Jackson, D. Appelt, J. Bear, R. Moore, and A. Podlozny, "A template matcher for robust NL interpretation," in *Proc. DARPA Speech and Natural Language Workshop*, 1991, pp. 190–194.

[43] D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler, and N. Morgan, "The Berkeley restaurant project," in *Proc. ICSLP*, 1994, pp. 2139–2142.

[44] A. Kellner, B. Rueber, and H. Schramm, "Using combined decisions and confidence measures for name recognition in automatic directory assistance systems," in *Proc. ICSLP*, 1998, pp. 2859–2862.

[45] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech*, 1999, pp. 2725–2728.
[46] D. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, 1987.
[47] J. Kowtko and P. Price, "Data collection and analysis in the air travel planning domain," in *Proc. DARPA Speech and Natural Language Workshop*, Oct. 1989.
[48] L. Lamel, S. Bennacef, J. L. Gauvain, H. Dartigues, and J. Temem, "User evaluation of the mask kiosk," in *Proc. ICSLP*, 1998, pp. 2875–2878.
[49] L. Lamel, S. Rosset, J. L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts, "The LIMSI ARISE system," in *Proc. IVTTA*, 1998, pp. 209–214.
[50] R. Lau, G. Flammia, C. Pao, and V. Zue, "WebGalaxy—Integrating spoken language and hypertext navigation," in *Proc. Eurospeech*, 1997, pp. 883–886.
[51] E. Levin, R. Pieraccini, and W. Eckert, "Using Markov decision process for learning dialogue strategies," in *Proc. ICASSP*, 1998, pp. 201–204.
[52] "Speech perception by humans and machines," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997.
[53] K. Livescu, "Analysis and modeling of nonnative speech for automatic speech recognition," S.M. thesis, MIT, Cambridge, MA, 1999.
[54] S. Marcus *et al.*, "Prompt constrained natural language—Evolving the next generation of telephony services," in *Proc. ICSLP*, 1996, pp. 857–860.
[55] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
[56] D. McDonald and L. Bolc, Eds., *Natural Language Generation Systems (Symbolic Computation Artificial Intelligence)*. Berlin, Germany: Springer-Verlag, 1998.
[57] K. McKeown, S. Pan, J. Shaw, D. Jordan, and B. Allen, "Language generation for multimedia healthcare briefings," presented at the *Proc. Applied Natural Language*, 1997.
[58] H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue, "A conversational system in the automobile classifieds domain," in *Proc. ICSLP*, 1996, pp. 542–545.
[59] S. Miller, R. Schwartz, R. Bobrow, and R. Ingria, "Statistical language processing using hidden understanding models," in *Proc. ARPA Speech and Natural Language Workshop*, 1994, pp. 278–282.
[60] R. Moore, D. Appelt, J. Dowding, J. Gawron, and D. Moran, "Combining linguistic and statistical knowledge sources in natural-language processing for ATIS," in *Proc. ARPA Spoken Language Systems Workshop*, 1995, pp. 261–264.
[61] H. Noguchi and Y. Den, "Prosody-based detection of the context of backchannel responses," in *Proc. ICSLP*, 1998, pp. 487–490.
[62] E. Nöth, "On the use of prosody in automatic dialogue understanding," in *Proc. ESCA Workshop Dialogue and Prosody*, 1999, pp. 25–34.
[63] A. Oh, "Stochastic natural language generation for spoken dialog systems," M.S. thesis, Carnegie-Mellon Univ., Pittsburgh, PA, May 2000.
[64] M. Ostendorf, C. Wightman, and N. Veilleux, "Parse scoring with prosodic information: An analysis/synthesis approach," *Comput. Speech Lang.*, vol. 7, no. 3, pp. 193–210, 1993.
[65] S. Oviatt, "Multimodal interfaces for dynamic interactive maps," in *Proc. Conf. Human Factors in Computing Systems: CHI'96*, 1996, pp. 95–102.
[66] D. Pallett, J. Fiscus, W. Fisher, J. Garafolo, B. Lund, A. Martin, and M. Pryzbocki, "Benchmark tests for the ARPA spoken language program," in *Proc. ARPA Spoken Language Systems Technology Workshop*, 1995, pp. 5–36.
[67] C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding systems," in *Proc. ICSLP*, 1998, pp. 815–818.
[68] K. Papineni, S. Roukos, and R. Ward, "Maximum likelihood and discriminative training of direct translation models," in *Proc. ICASSP*, 1998, pp. 189–192.
[69] J. Peckham, "A new generation of spoken dialogue systems: Results and lessons from the SUNDIAL project," in *Proc. Eurospeech*, 1993, pp. 33–40.
[70] R. Pieraccini and E. Levin, "Stochastic representation of semantic structure for speech understanding," in *Speech Commun.*, vol. 11, 1992, pp. 283–288.
[71] R. Pieraccini, E. Levin, and W. Eckert, "AMICA: The AT&T mixed initiative conversational architecture," in *Proc. Eurospeech*, 1997, pp. 1875–1879.
[72] J. Polifroni, S. Seneff, J. Glass, and T. Hazen, "Evaluation methodology for a telephone-based conversational system," in *Proc. Int. Conf. on Lang. Resources and Evaluation*, 1998, pp. 42–50.
[73] J. Polifroni and S. Seneff, "Galaxy-II as an architecture for spoken dialogue evaluation," in *Proc. Int. Conf. Language. Resources and Evaluation*, Athens, Greece, May 31–June 2, 2000.
[74] P. Price, "Evaluation of spoken language systems: The ATIS domain," in *Proc. DARPA Speech and Natural Language Workshop*, 1990, pp. 91–95.
[75] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. Cambridge, MA: Cambridge Univ. Press, 2000.
[76] S. Rosset, S. Bennacef, and L. Lamel, "Design strategies for spoken language dialog systems," in *Proc. Eurospeech*, 1999, pp. 1535–1538.
[77] A. Rudnicky, M. Sakamoto, and J. Polifroni, "Evaluating spoken language interaction," in *Proc. DARPA Speech and Natural Language Workshop*, Oct. 1989, pp. 150–159.
[78] A. Rudnicky, J. M. Lunati, and A. Franz, "Spoken language recognition in an office management domain," in *Proc. ICASSP*, 1991, pp. 829–832.
[79] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh, "Creating natural dialogs in the carnegie mellon communicator system," in *Proc. Eurospeech*, 1999, pp. 1531–1534.
[80] D. Sadek, "Design considerations on dialogue systems: From theory to technology—The case of Artimis," in *Proc. ESCA Workshop Interactive Dialogue in Multi-Modal Systems*: Kloster Irsee, 1999, pp. 173–188.
[81] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR ν-talk speech synthesis system," in *Proc. ICSLP*, 1992, pp. 483–486.
[82] A. Sanderman, J. Sturm, E. den Os, L. Boves, and A. Cremers, "Evaluation of the dutch train timetable information system developed in the ARISE Project," in *Proc. IVTTA*, 1998, pp. 91–96.
[83] J. van Santen, L. Pols, M. Abe, D. Kahn, E. Keller, and J. Vonwiller, "Report on the 3rd ESCA TTS workshop evaluation procedure," in *Proc. 3rd ESCA Workshop Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 329–332.
[84] E. Shriberg *et al.*, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Language Speech*, vol. 41, pp. 439–447, 1998.
[85] S. Seneff, "TINA: A natural language system for spoken language applications," *Comput. Linguistics*, vol. 18, no. 1, pp. 61–86, 1992.
[86] ——, "Robust parsing for spoken language systems," in *Proc. ICASSP*, 1992, pp. 189–192.
[87] S. Seneff, V. Zue, J. Polifroni, C. Pao, L. Hetherington, D. Goddeau, and J. Glass, "The preliminary development of a displayless PEGASUS system," in *Proc. ARPA Spoken Language Technology Workshop*, 1995, pp. 212–217.
[88] S. Seneff, D. Goddeau, C. Pao, and J. Polifroni, "Multimodal discourse modeling in a multi-user multi-domain environment," in *Proc. ICSLP*, 1996, pp. 188–191.
[89] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "GALAXY-II: A reference architecture for conversational system development," in *Proc. ICSLP*, 1998, pp. 931–934.
[90] S. Seneff, R. Lau, and J. Polifroni, "Organization, communication, and control in the GALAXY-II conversational system," in *Proc. Eurospeech*, 1999, pp. 1271–1274.
[91] S. Seneff, R. Lau, J. Glass, and J. Polifroni, "The mercury system for flight browsing and pricing," *MIT Spoken Language System Group Annual Progress Rep.*, pp. 23–28, 1999.
[92] S. Seneff and J. Polifroni, "A new restaurant guide conversational system: Issues in rapid prototyping for specialized domain," in *Proc. ICSLP*, 1996, pp. 665–668.
[93] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, "SABLE: A standard for TTS Markup," in *Proc. ICSLP*, 1998, pp. 1719–1722.
[94] D. Stallard and R. Bobrow, "Fragment processing in the DELPHI system," in *Proc. DARPA Speech and Natural Language Workshop*, 1992, pp. 305–310.
[95] V. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant: Strategies for spoken dialogue systems," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 51–62, Jan. 2000.

[96] J. Sturm, E. den Os, and L. Boves, "Dialogue management in the Dutch ARISE train timetable information system," in *Proc. Eurospeech*, 1999, pp. 1419–1422.
[97] S. Sutton, E. Kaiser, A. Cronk, and R. Cole, "Bringing spoken language systems to the classroom," in *Proc Eurospeech*, 1997, pp. 709–712.
[98] S. Sutton *et al.*, "Universal speech tools: The CSLU toolkit," in *Proc. ICSLP*, 1998, pp. 3221–3224.
[99] D. Thomson and J. Wisowaty, "User confusion in natural language services," in *Proc. ESCA Workshop Interactive Dialogue in Multi-Modal Systems*: Kloster Irsee, 1999, pp. 189–196.
[100] M. Walker, D. Litman, C. Kamm, and A. Abella, "PARADISE: A general framework for evaluating spoken dialogue agents," in *Proc. ACL/EACL*, 1997, pp. 271–280.
[101] M. Walker, J. Boland, and C. Kamm, "The utility of elapsed time as a usability metric for spoken dialogue systems," in *Proc. ASRU Workshop*, Keystone, CO, 1999, pp. 1167–1170.
[102] W. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue, "Yinhe: A Mandarin Chinese version of the galaxy system," in *Proc. Eurospeech*, 1997, pp. 351–354.
[103] N. Ward, "Using prosodic cues to decide when to produce back-channel utterances," in *Proc. ICSLP*, 1996, pp. 1728–1731.
[104] W. Ward, "Modeling nonverbal sounds for speech recognition," in *Proc. DARPA Workshop Speech and Natural Language*, 1989, pp. 47–50.
[105] ——, "The CMU air travel information service: Understanding spontaneous speech," in *Proc. ARPA Workshop Speech and Natural Language*, 1990, pp. 127–129.
[106] ——, "Integrating semantic constraints into the SPHINX-II recognition search," in *Proc. ICASSP*, 1994, pp. II-17–II–20.
[107] W. Ward and S. Issar, "Recent improvements in the CMU spoken language understanding system," in *Proc. ARPA Human Language Technology Workshop*, 1996, pp. 213–216.
[108] J. Yi and J. Glass, "Natural-sounding speech synthesis using variable length units," in *Proc. ICSLP*, 1998, pp. 1167–1170.
[109] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill, "PEGASUS: A spoken language interface for on-line air travel planning," *Speech Commun.*, vol. 15, pp. 331–340, 1994.
[110] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 1, pp. 85–96, 2000.

Victor W. Zue received the Dr.(sci.) degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1976.

He is currently a Senior Research Scientist at MIT, Associate Director of the Institute's Laboratory for Computer Science, and Head of its Spoken Language Systems Group. His main research interest is in the development of conversational interfaces to facilitate graceful human/computer interactions. He has taught many courses at MIT and around the world, and he has also written extensively on this subject.

Dr. Zue is a Fellow of the Acoustical Society of America. In 1994, he was elected Distinguished Lecturer by the IEEE Signal Processing Society.

James R. Glass (Member, IEEE) received the B.Eng. degree from Carleton University, Ottawa, Canada, in 1982 and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1984, and 1988, respectively.

Since then, he has worked in MIT's Laboratory for Computer Science, where he is currently a Principal Research Scientist and Associate Head of the Spoken Language Systems Group. His research interests include acoustic-phonetic modeling, speech recognition and understanding, and corpus-based speech synthesis.

Dr. Glass served as a member of the IEEE Acoustics, Speech, and Signal Processing Speech Technical Committee from 1992 to 1995. From 1997 to 1999, he served as an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.