Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10017–10022, October 1995
Colloquium Paper

This paper was presented 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 Irvine, CA, February 8–9, 1993.

Deployment of human–machine dialogue systems

DAVID B. ROE

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT **The deployment of systems for human-to-machine communication by voice requires overcoming a variety of obstacles that affect the speech-processing technologies. Problems encountered in the field might include variation in speaking style, acoustic noise, ambiguity of language, or confusion on the part of the speaker. The diversity of these practical problems encountered in the "real world" leads to the perceived gap between laboratory and "real-world" performance. To answer the question "What applications can speech technology support today?" the concept of the "degree of difficulty" of an application is introduced. The degree of difficulty depends not only on the demands placed on the speech recognition and speech synthesis technologies but also on the expectations of the user of the system. Experience has shown that deployment of effective speech communication systems requires an iterative process. This paper discusses general deployment principles, which are illustrated by several examples of human–machine communication systems.**

Speech-processing technology is now at the point at which people can engage in voice dialogues with machines, at least in limited ways. Simple voice communication with machines is now deployed in personal computers, in the automation of long-distance calls, and in voice dialing of mobile telephones. These systems have small vocabularies and strictly circumscribed task domains. In research laboratories there are advanced human–machine dialogue systems with vocabularies of thousands of words and intelligence to carry on a conversation on specific topics. Despite these successes, it is clear that the truly intelligent systems envisioned in science fiction are still far in the future, given the state of the art today.

Human–machine dialogue systems can be represented as a four-step process, as shown in Fig. 1. This figure encompasses both the simple systems deployed today and the spoken language understanding we envision for the future. First, a speech recognizer transcribes sentences spoken by a person into written text (1, 2). Second, a language understanding module extracts the meaning from the text (3, 4). Third, a computer (consisting of a processor and a database) performs some action based on the meaning of what was said. Fourth, the person receives feedback from the computer in the form of a voice created by a speech synthesizer (5, 6). The boundaries between these stages of a dialogue system may not be distinct in practice. For instance, language-understanding modules may have to cope with errors in the text from the speech recognizer, and the speech recognizer may make use of grammar and semantic constraints from the language module in order to reduce recognition errors.

In the 1993 "Colloquium on Human–Machine Communication by Voice," sponsored by the National Academy of Sciences (NAS), much of the discussion focused on practical difficulties in building and deploying systems for carrying on voice dialogues between humans and machines. Deployment of systems for human-to-machine communication by voice requires solutions to many types of problems that affect the speech-processing technologies. Problems encountered in the field might include variation in speaking style, noise, ambiguity of language, or confusion on the part of the speaker. There was a consensus at the colloquium that a gap exists between performance in the laboratory and accuracy in the field, because conditions in real applications are more difficult. However, there was little agreement about the cause of this gap in performance or what to do about it.

A key point of discussion at the NAS colloquium concerned the factors that make a dialogue easy or difficult. Many such degrees of difficulty were mentioned in a qualitative way. To summarize the discussion in this paper it seems useful to introduce a more formal concept of the degree of difficulty of a human–machine dialogue and to list each dimension that contributes to the overall difficulty. The degree of difficulty is a useful concept, despite the fact that it is only a "fuzzy" (or qualitative) measure because of lack of precision in quantifying an overall degree of difficulty for an application.

A second point of discussion during the NAS colloquium concerned the process of deployment of human–machine dialogue systems. In several cases such systems were built and then modified substantially as the designers gained experience in what the technology could support or in user-interface issues. This paper elaborates on this iterative deployment process and contrasts it with the deployment process of more mature technologies.

DEGREE OF DIFFICULTY OF A VOICE DIALOGUE APPLICATION

Whether a voice dialogue system is successful depends on the difficulty of each of the four steps in Fig. 1 for the particular application, as well as the technical capabilities of the computer system. There are several factors that can make each of these four steps difficult. Unfortunately, it is difficult to quantify precisely the difficulty of these factors. If the technology performs unsatisfactorily at any stage of processing of the speech dialogue, the entire dialogue will be unsatisfactory. We hope that technology will eventually improve to the point that there are no technical barriers whatsoever to complex speech-understanding systems. But until that time it is important to know what is easy, what is difficult but possible, and what is impossible, given today's technology.

What are the factors that determine the degree of difficulty of a voice dialogue system? In practice, there are several factors for each step of the voice dialogue that may make the task difficult or easy. Because these factors are qualitatively independent, they can be viewed as independent variables in a multidimensional space. For a simple example of two dimensions of difficulty for speech recognition refer to Fig. 2.

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.

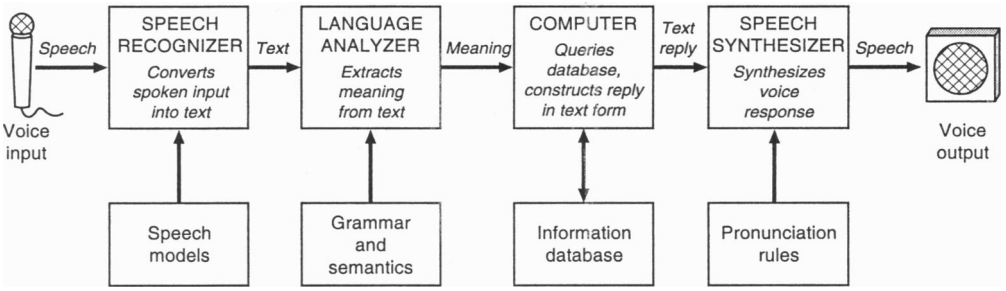

FIG. 1. A human–machine dialogue system.

There are many dimensions of difficulty for a speech recognition system, two of which are shown. Eight applications are rated according to difficulty of speaking mode (vertical axis) and vocabulary size (horizontal axis). "Voice dictation" refers to commercial 30,000-word voice typewriters; "V.R.C.P." stands for voice recognition call processing, as a way to automate long-distance calling; "telephone number dialing" refers to connected digit recognition of phone numbers; "DARPA resource management" refers to the 991-word Naval Resource Management task with a constraining grammar; "A.T.I.S." stands for the DARPA (Defense Advanced Research Projects Agency) Air Travel Information System; and "natural spoken language" refers to conversational speech on any and every topic. Clearly, a task that is difficult in both the dimensions of vocabulary size and speaking style would be harder (and would have lower accuracy) than a small, isolated word recognizer, if all other factors are equal. The other factors are not equal, as discussed in the section "Dimensions of the Recognition Task." Note that telephone number dialing, despite its position on the two axes in this figure, is a difficult application because of dimensions not shown here, such as user tolerance of errors and grammar perplexity.

Ideally, a potential human–machine dialogue could receive a numerical rating along each dimension of difficulty, and a cumulative degree of difficulty could be computed by summing the ratings along each separate dimension. Such a quantitative approach is overly simplistic. Nevertheless, it is a valuable exercise to evaluate potential applications qualitatively along each of the dimensions of difficulty.

The problems for voice dialogue systems can be separated into those of speech recognition, language understanding, and speech synthesis, as in Fig. 1. (For the database access stage, a conventional computer is adequate for most voice dialogue tasks. The data-processing capabilities of today's machines pose no barriers to development of human–machine communication systems.) Let us examine the steps of speech recognition, language understanding, and speech synthesis in order to analyze the specific factors, or dimensions of difficulty, that make an application easy or difficult.

Dimensions of the Speech Recognition Task

Humans are able to understand speech so readily that they often fail to appreciate the difficulties that this task poses for machines. The exception may be the process of learning a foreign language. Indeed, there is more than a casual relationship between the problems faced by an adult listening to a foreign language and those of a machine recognizing speech.

The performance of speech recognizers is typically assessed by measuring the accuracy of the recognizer, or equivalently, its error rate. But the accuracy of any recognizer may vary widely, depending on the conditions of the experiment and the speech data. John Makhoul, in his paper (1), has listed some rules of thumb indicating how recognition accuracies vary. In the laboratory, speech recognizers are quite accurate in acoustic pattern matching. In real-world conditions, however, the error rate is much higher, due in part to the increased variability of speech styles encountered. Given high-quality, consistent speech samples recorded in a quiet laboratory, and given sufficient speech samples to fully train an HMM (hidden Markov model), accuracies are almost comparable to human accuracies in acoustic perception. For instance, numbers can be recognized with an error rate of less than one in 300 words (99.7 percent accuracy) (8). This result was obtained on the Texas Instruments/National Institute of Standards and Technology speech database recorded under laboratory conditions in a soundproof booth and with a balance of dialects of native U.S. speakers. On a larger, speaker-independent task, the DARPA resource management task described below, word accuracies of about 96 percent can be achieved on a vocabulary of a thousand words (9). Again, these results are obtained by using carefully spoken speech recorded in a quiet environment.

In applications the variability of speech and speaking environments is much greater, so that the same speech recognition

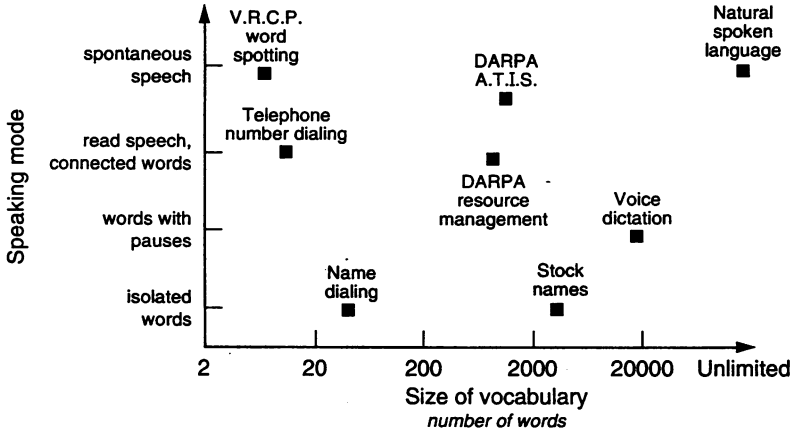

FIG. 2. Two (of many) dimensions of difficulty for speech recognition (adapted from ref. 7).

algorithm will have error rates that are much higher than with well-controlled laboratory speech. For instance, in tests of speech recognition of credit card numbers spoken by merchants at retail stores, the error rate rises to about 2 percent per digit (from 0.3 percent) (10) with an algorithm similar to that described by Gauvin and Lee (8). This increase in error rate is typical of the difference between laboratory-quality speech and the speech encountered in field conditions. In fact, speech recognition in the field is a harder task because the speech is more variable than laboratory speech.

The following are the dimensions of difficulty for speech recognition applications:

- *Speaker independence*. It is much easier to characterize an individual speaker's voice than to recognize all voice types and all dialects. Applications can be categorized, in increasing order of difficulty, as speaker-trained, speaker-adaptive, multispeaker, speaker-independent, and speaker-independent with nonnative speakers.
- *Expertise of the speaker*. People typically learn how to get good recognition results with practice. Applications in which the speakers can quickly gain experience are more likely to succeed than applications in which the majority of people use the service infrequently. As Richard Schwartz remarked at the NAS colloquium, "You are a first time user only once."
- *Vocabulary confusability*. Other things being equal, a larger vocabulary is more likely to contain confusable words or phrases that can lead to recognition errors. However, some small vocabularies may be highly confusable. The letters of the alphabet ("A, B, C, D. . .") are notoriously difficult to recognize.
- *Grammar perplexity*. An application may have a grammar in which only certain words are permitted given the preceding words in a sentence, which reduces the opportunity for errors. The perplexity of the grammar is the average number of choices at any point in the sentence.
- *Speaking mode: rate and coarticulation*. Speech sounds are strongly affected by surrounding sounds in rapid speech. Isolated words are more consistently recognized than words in fluent speech. Voice dictation systems typically require that speakers leave a slight pause between words in a sentence. Speaker variabilities, including disfluencies such as hesitation, filled pauses, and stammering, are also important. People are able to understand these normal variations in speed or loudness and to compensate for any involuntary changes caused by stress upon the speaker. Only in the simplest cases can machines handle such conditions.
- *Channel conditions*. Speech that is distorted or obscured by noise is more difficult for machines to recognize than high-quality speech. Noise can include background speech and other acoustic noise as well as noise in the transmission channel. Variability in transmission bandwidth and in microphone characteristics also affects speech recognition accuracy.
- *User tolerance of errors*. It is important to bear in mind that voice dialogue systems, notwithstanding recent advances, remain error-prone. Given that any speech recognizer will make occasional errors, the inconvenience to the user should be minimized. This means that careful design of human factors of an application will be essential. The central questions when considering an application using a speech recognizer are: (*i*) What accuracy will the user of this service expect? (*ii*) Is the speech recognizer accurate enough to meet the expectations of the user? (*iii*) Does the benefit of using speech recognition in this application outweigh its cost, compared to alternative technologies? Each of these dimensions of difficulty embodies some aspect of speech variability, which is the central problem of speech recognition. The more sophisticated the speech recognizer, the better it is able to cope with these practical difficulties. Increasing the robustness of speech recognizers to all types of variability is a major challenge of current speech recognition research. These sources of variability must be carefully considered when planning applications of the technology, because it is these robustness characteristics that determine whether a speech recognizer will be accurate enough to be satisfactory to the users.

Dimensions of the Language-Understanding Task

The difficulties of natural language understanding are well known (3, 11, 12) and will not be discussed in detail here. For speech-understanding systems the difficulties are compounded by uncertainty in interpreting the acoustic signal, which results in errors in the text (4). Therefore, spoken-language-understanding systems are now limited to constrained domains in which there is little ambiguity. Furthermore, models used for speech dialogue systems tend to be simpler and less powerful than those used for text understanding. Though it is widely known that finite-state grammars are too simple to express the range of meanings of English, these elementary grammars are typically used by speech-understanding systems. For instance, voice dictation systems by IBM (13), Dragon Systems, and Kurzweil Applied Intelligence use simple *N*-gram models that estimate the probability of sequences of up to three words based on training texts. The dimensions of difficulty of language understanding are:

- *Grammar complexity and ambiguity*. The wider the range of meanings in the domain of understanding, the more complex the grammar must be to express those meanings. This complexity leads to a greater possibility of semantic ambiguity, that is, the chance that an input text sequence may have more than one possible interpretation. Finally, semantic or grammatical ambiguity may be compounded by acoustic ambiguity (words that sound similar) or speech recognition errors.
- *Language variability*. Language is very flexible. For any meaning there are many ways of expressing it. As a trivial example, there are well over 50 ways of saying "no" in English, ranging from "probably not" to "over my dead body." The degree to which the user chooses an unusual phrasing creates problems for a speech-understanding system.
- *Rejection of "off-the-subject" input*. In some applications the users may respond with a reasonable sentence that is beyond the scope of the system's language model. A collect-call system that is supposed to understand "yes" and "no" may have a difficult time coping with the response "I'll get Mommy." In cases like these the system may misrecognize the response and take some wildly incorrect action because of its misunderstanding. In some applications it is possible to train the users to avoid sentences that the system cannot understand, but this is not always practical. How can the system recognize what it does not "know"? Lynn Bates, in a comment at the NAS colloquium, has suggested building a separate language model just to catch the most frequent user responses that are out of the range of the original, more constrained language model. This second language model could be used to help prompt the user on what to say to be understood by the machine.

More powerful statistical techniques now being developed for text understanding (14) hold the promise of significantly improving the language understanding capabilities of advanced voice dialogue systems. Alex Waibel remarked at the colloquium that it is very time consuming to build special-purpose speech-understanding systems and that the long-term goal should be to create a machine that could learn the specific application with repeated practice. With self-organizing systems such as neural networks, it might someday be possible to build this type of learning system. Recalling Richard Schwartz's comment about people becoming experts in using voice dialogue systems, it might be more practical in the short term to provide people with the feedback they need to adapt to the system rather than expect machines to adapt to people.

Dimensions of the Speech Synthesis Task

There are two families of computer speech technologies today: digitized human speech and text-to-speech synthesis. Text-to-speech synthesis is flexible enough to pronounce any sentence but lacks the naturalness of recorded human speech. Digitized human speech is natural sounding but inflexible because only prerecorded phrases can be spoken. Text-to-speech systems are able to synthesize any text with an intelligibility almost as high as a human speaker. However, it is a major challenge to achieve naturalness in synthesized speech (5, 6).

For a speech output application the dimensions of difficulty relate to problems in synthesizing intelligible and pleasant-sounding computer speech, as opposed to speech understanding. The dimensions of difficulty are as follows:

- *Quantity of text*. It is impractical to record huge amounts of human speech (say, more than 100 hours) for a speech dialogue system. The vast majority of current applications in the voice response industry use recorded human speech. With digitized human speech, or waveform coding, the quality is limited only by the skill of the speaker and by the compression algorithm for the recorded speech, typically 2000 to 8000 bytes per second of speech. Recorded human speech has a major drawback, however, because every sentence must be recorded separately. Splicing together a phrase out of individually recorded words is unsatisfactory because of the "choppy" quality of the concatenated sentence. For applications in which a great variety of sentences must be spoken, or one in which the information changes frequently so that recording is impractical, text-to-speech synthesis must be used.
- *Variability of the input text*. There are applications in which the text being processed may contain abbreviations, jargon, or outright errors. Playing back electronic mail messages is one such application that must cope with error-prone input text, whereas pronouncing the names of catalog items has low variability. Also, specialized text preprocessors can be created for pronunciation of special vocabularies (such as prescription drug names); this is not practical for coverage of unrestricted English. When the text has low variability and consists of a few fixed phrases, recorded human speech can be used.
- *Length of the sentences and grammatical complexity*. Longer sentences tend to have more grammatical and semantic structure than short phrases, and current text-to-speech synthesizers provide only rudimentary linguistic analysis of the text (5). Therefore, there is a tendency for longer, more complex sentences to have a poorer subjective rating than short simple phrases (15). An application in which the text is very complex, such as reading Shakespearean sonnets, would be more difficult than pronouncing words or short phrases.
- *Expectations of the listener*. Listeners are likely to be tolerant of a technology that provides them with a new and valuable service but intolerant of a system of poorer quality than what they are used to. For instance, consumers reacted positively to a service known as "Who's Calling?" in which they heard a text-to-speech synthesizer say "you have a call from the phone of John Doe" because this is a service not available before. But in other applications the quality of text-to-speech synthesis is a concern. In subjective tests of speech quality, text-to-speech synthesizers are judged significantly worse than digitized human speech (15). This remains true even when the text-to-speech synthesizer was provided with the pitch contour and the phoneme durations used by the original human speaker.

Intelligibility for text-to-speech systems has been an issue in the past, but the intelligibility of modern systems is high enough for most applications. Word intelligibility measured at the word level is only slightly lower for good text-to-speech systems than for digitized human speech (15). However, intelligibility at the sentence level can be impaired when the prosody of complex sentences is so mangled that the meaning is obscured.

Additional Dimensions of Difficulty

In addition to the dimensions of difficulty based on the limitations of the technologies of speech processing, there are engineering constraints on the deployment of any system: cost, size, power, and time available for deployment. In particular, cost, size, and power affect the amount of memory available for the speech processing and the power of the speech-processing chips. Hand-held devices and consumer equipment have particularly severe constraints.

- *Memory and processor requirements*. Speech recognition algorithms require millions of operations per second. Considerable ingenuity has been exerted to implement algorithms efficiently. In order of decreasing cost and computation power, recognizers have been programmed on parallel processors, RISC chips, floating point digital signal processors, general-purpose microprocessors, and integer digital signal processors. For speech synthesis, waveform coding systems require little processing power (a small fraction of the processing power of a digital signal processor chip), and the memory is proportional to the amount of speech to be played back. On the other hand, text to speech requires a high-speed microprocessor and between 0.5 and 5 megabytes of memory. For hand-held pronouncing dictionaries, text-to-speech synthesis is used because it is too costly to provide memory to store the waveform for each word in the dictionary.
- *System integration requirements*. As with other technologies, applications that rely on networked databases or processors, that need to access information for multiple users, that must coexist with existing equipment, or that must be compatible with future products and services will require more care during the systems engineering process than do stand-alone applications.

Examples of Speech Applications

We have listed some of the dimensions of difficulty for a human–machine voice dialogue system. In principle, one might be able to rate an application along each dimension as "easy" or "difficult," thus arriving at an overall degree of difficulty of the application. Actual voice dialogue systems are not so easily quantified. But clearly an application that is rated "difficult" along most of the dimensions will require extraordinary effort to deploy. Table 1 shows four human–machine communication systems with voice input/output.

1. VRCP (Voice Recognition Call Processing) is an AT&T service that automates the operator's role in placing long-distance calls (16).
2. The DARPA ATIS task is an experimental system for getting information from a database of airline flights (9).
3. Voice dialing refers to cellular telephones with speech recognition capability, so that calls may be made or received while driving a car without dialing by hand. Cellular telephones with voice-dialing are sold by AT&T, Motorola, Nippon Electric Co., and others.
4. StockTalk (17) is a system trialed recently by Bell Northern Research that allows people to get current prices of securities on several stock exchanges by speaking the name of the company.

Table 1 evaluates these four applications along each of the dimensions of difficulty for speech recognition, language understanding, and speech synthesis. The message is that each of these applications has some dimensions that are difficult and some that are easy. These are all cutting-edge systems in their own way, but the dimensions in which they excel are different.

Table 1. Degree of difficulty for four voice applications

Dimension of difficulty	Application			
	AT&T's VRCP	DARPA ATIS	Cellular phone with name dialing	Bell Northern Research's stock talk
Speaker independence	Difficult: various dialects and nonnative speakers	Moderate: mostly native American English speakers	Easy: trained to one speaker	Difficult: many dialects and nonnative speakers
Speaker expertise	Difficult: high proportion of first-time users	Moderate: speakers have time to rehearse a query	Easy: the owner is trained by the telephone	Moderate: First-time users have opportunity to practice
Vocabulary size (acoustic confusibility)	Easy: Seven dissimilar words	Difficult: unlimited vocabulary with confusible words	Moderate: user may train similar-sounding names	Moderate: over 2000 words, but most are dissimilar
Grammar perplexity (number of choices)	Easy	Moderate: perplexity approx. 50, but difficult to specify grammar	Moderate: perplexity approx. 60	Difficult: no grammar; high perplexity
Speaking mode (coarticulation and disfluencies)	Difficult: continuous extraneous speech, with barge in	Moderate: continuous speech with some disfluencies	Easy: isolated words	Easy: isolated words
Channel variability (acoustic or electrical)	Moderate: telephone channel with handset	Easy: quiet laboratory conditions	Difficult: high noise levels, mike far from mouth	Moderate: telephone channel with handset
User tolerance of errors	Moderate: a human operator is always available	N.A.	Moderate: user can hang up before a call is placed incorrectly	Easy: very little penalty for incorrect recognition
Grammar complexity and ambiguity	Easy	Difficult: many complex sentences; context required	Easy	Easy: no grammar
Language variability	Moderate: synonyms for "yes" and "no"	Difficult: wide variety of sentence forms	Easy: though users forget how they trained names	Moderate: some stocks have several possible names
Rejection of extraneous speech	Moderate: must reject casual speech, answering machines	Easy: all "incorrect" queries excluded	Difficult: microphone is on at all times	N.A., no rejection
Quantity of speech to synthesize	Easy: uses prerecorded speech prompts	Difficult: TTS must synthesize millions of sentences	Easy: records user's voice for name feedback	Moderate: TTS used for company names
Variability of input text	N.A.	Moderate: text in the database can be prescreened	N.A.	Easy: pronounciations can be verified in advance
Length of sentence and grammatical complexity	N.A.	Difficult: long, complex sentences	N.A.	Easy: short, structured phrases
Listener expectations	High: recorded speech	Moderate	Moderate: should resemble user's voice	Easy: should be intelligible
Processor and memory requirements	Moderate: multichannel system	Easy: must run in close to real time on a workstation	Difficult: low-cost, single-chip processor	Moderate: multi-channel system

N.A., not applicable.

PROCEDURE FOR DEPLOYMENT OF SPEECH APPLICATIONS

The design and development of a voice transaction system is an iterative one. While one might think that a designer specifies the voice transaction, builds the speech input and output modules, and deploys the final system, this is not the case in practice. Because of the complexity and variety of dialogues, it is exceedingly difficult to anticipate all the factors that are critical to success. Experience has shown that developing a human–machine dialogue system is an iterative process.

A typical iterative process is shown in Fig. 3. In the initial design the overall objectives of the system are set up. Development proceeds in a conventional way until the trial system is set up in field conditions. The system designers need to have an auditing system in place to determine what people say during voice transactions and what the machine's response is. Specific problems will be identified and diagnosed to correct the system's accuracy. Human–machine dialogue systems are different from more mature technologies in that they require several iterations.

With mature technologies it is possible to schedule a development timetable with a fair degree of confidence that the technology will work as planned and on schedule. Rarely if ever has this been possible with voice dialogue systems. The design of a user interface (18) is painstaking because people respond in unexpected ways. For instance, in automation of operator services, people were asked to say one of five phrases: "collect," "calling card," "operator," "person to person," and "third number." Twenty percent of the callers spoke these phrases with other words such as "please" or "uhhm." Rewording the voice prompts could reduce this problem but only with an unacceptably longer-duration prompt. The preferred solution, word spotting in speech recognition, took several years to develop and deploy. Second, it is difficult to gather speech for training speech recognizers unless you have a

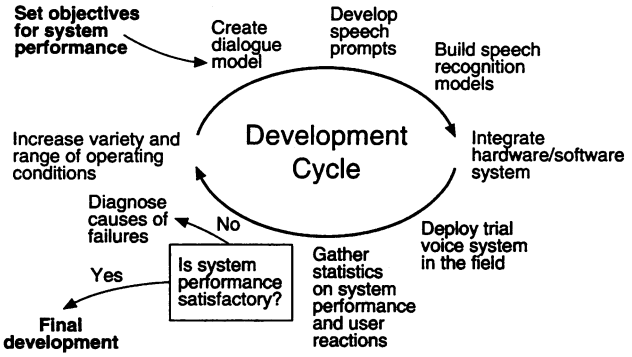

FIG. 3. Deployment process for a voice dialogue system.

working system that will capture speech in exactly the environment encountered in the real service. Therefore, the speech recognition accuracy will be lower than expected based on the style of speech used for training.

The procedure outlined above for deployment of speech technology may seem ad hoc, but it is necessary given the current maturity of the technology. When an engineer designs a bridge, there is an initial design, there are calculations to check the structural soundness, the bridge is constructed, and the bridge functions as it was designed to. It is not necessary to build a "trial" bridge to find out why it is going to fall down or to design the "final" bridge by successive approximation. In this respect, speech technology is still immature. In some sense we still lack a complete set of design principles needed to guarantee the integrity of a human–machine dialogue system.

The Art of Human–Machine Dialogues

Current voice dialogue practice encompasses engineering art as well as scientific knowledge. Fundamental knowledge of speech production and basic principles of pattern matching have been essential to the success of speech recognition over the past 25 years. That said, the art of successful engineering is critically important for applications of voice dialogue systems (19). There is an important element of craftsmanship in building a successful speech transaction. Often, this engineering art has been gained through trial and error. It should be emphasized that improving the engineering art is a proper and necessary topic for applied research.

The engineering art of speech recognition has improved significantly in the past few years, further opening up the range of possible applications.

- *Subword units*. It is now possible to build a dictionary of models comprised of constituent phonetic (or phoneme-like) statistical models, first for small, easily distinguishable vocabularies, and later for larger vocabularies. The effort and expense of gathering speech from many speakers for each vocabulary word have been reduced.
- Noise immunity. Better speech enhancement algorithms and models of background noise make speech recognizers more accurate in noisy or changing environments, such as automobiles.
- *Speaker adaptation*. People can adapt quickly to dialects and accents in speech. Machines now have the beginnings of the capability to respond more accurately as they learn an individual voice.
- *Rudimentary language understanding*. The ability to spot key words in a phrase is the first step toward understanding the essence of a sentence even if some words are not recognized.
- *"Barge in."* It is sometimes desirable, when talking with a person, to be able to interrupt the conversation. In telephone-based voice response systems, it is possible to interrupt a prompt using Touch-Tones. This capability has been extended to allow users the ability to speak during a prompt and have the system recognize them.
- *Rejection*. An ability that people take for granted in conversation is the ability to detect when we do not understand. Unfortunately, this is a most difficult task for current speech recognition systems. While it is possible to determine when there are two (or more) possible words or sentences, it has been very difficult for systems to determine when people are saying something on a completely different subject. This can lead to comical, if not frustrating, results for the user. Further research is needed in detecting this type of "none of the above" response.

The design of an easy-to-use dialogue with a computer system is a significant challenge. We know from experience that is possible to design good human interfaces for computer dialogue systems. Unfortunately, it has also been verified that it is possible to design systems that aggravate people. At this time there are some general guidelines for good human interface design, but there is no "cookbook" recipe that guarantees a pleasant and easy-to-use system (18).

CONCLUSIONS

The concept of the degree of difficulty of a human–machine voice dialogue system can be used to evaluate its feasibility. The degree of difficulty of a particular application depends on many factors. Some are obvious, but others are easy to overlook. For example, the expertise of the users has a dramatic effect on the performance of these systems. Also, the willingness of users to overlook deficiencies in the system varies widely depending on whether there are other alternatives. A comprehensive view of all the dimensions of difficulty is needed in order to assess the overall degree of difficulty.

Deployment of voice transaction services is an iterative process. Because the machine must cope with errors made by the person, and the human being must cope with errors made by the machine, the nature of the transaction is difficult if not impossible to predict in advance. Though the ultimate goal is to create a machine that can adapt to the transaction as it gains more experience, the human–machine dialogue systems of today require engineering art as well as scientific principles.

1. Makhoul, J. & Schwartz, R. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9956–9963.
2. Rabiner, L. & Juang, B. H. (1993) *Fundamentals of Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).
3. Bates, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9977–9982.
4. Moore, R. C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9983–9988.
5. Allen, J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9946–9952.
6. Carlson, R. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9932–9937.
7. Atal, B. S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10046–10051.
8. Gauvin, J. & Lee, C. H. (1992) *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (San Francisco), pp. 481–484.
9. Marcus, M., ed. (1992) *Proceedings, Speech and Natural Language Workshop,* Harriman, NY (Kaufmann, San Mateo, CA).
10. Ramesh, P., *et al.* (1992) *Speech Commun.* **11,** 229–235.
11. Berwick, R. (1987) in *AI in the 1980's and Beyond*, eds. Grimson, W. E. & Patil, R. S. (Mass. Inst. of Technol. Press, Cambridge, MA), pp. 156–183.
12. Hirst, G. (1987) *Semantic Interpretation and the Resolution of Ambiguity* (Cambridge Univ. Press, Cambridge, U.K.).
13. Bahl, L. R., *et al.* (1989) in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Glasgow, Scotland), pp. 465–468.
14. Marcus, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10052–10059.
15. Van Santen, J. P. H. (1993) *Comput. Speech Lang.* **7,** 49–100.
16. Wilpon, J. G. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9991–9998.
17. Lennig, M., Sharp, D., Kenny, P., Gupta, V. & Precoda, K. (1992) in *Proceedings of the 1992 International Conference on Spoken Language Processing* (Banff, AB Canada), pp. 93–96.
18. Kamm, C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10031–10037.
19. Nakatsu, R. & Suzuki, Y. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10023–10030.