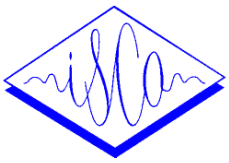

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

IDENTIFICATION OF PRINCIPAL ERGONOMIC REQUIREMENTS FOR INTERACTIVE SPOKEN LANGUAGE SYSTEMS

Stephen Springer, Sara Basson, and Judith Spitz

Speech Technology Group, Artificial Intelligence Laboratory
NYNEX Science & Technology, Inc.
500 Westchester Ave., White Plains, NY 10604, U.S.A.

ABSTRACT

Successful deployment of speech understanding systems demands an understanding of all relevant aspects of human/machine interaction. By determining the factors which most significantly affect user behavior, and quantitatively describing the effects of varying these factors, we can more accurately control and predict a system's ultimate performance under real user conditions. This paper describes the design of an automated system, *Money Talks*, which will determine speech recognition and application requirements in a real user, service-providing environment: a telephone-based information service. We discuss the reasoning and implications underlying our design decisions to investigate certain aspects of interaction, while not investigating others. We also discuss the trade-offs between "naturalness," efficiency, and the effect on speaker compliance associated with wording variations in prompts. Lastly, issues related to the design of such trials, such as the need for barge-in technology, realistic simulations of graceful failure, and operator-handoff procedures, are also addressed.

10.21437/ICSLP.1992-373

1. INTRODUCTION AND MOTIVATION

Cooperative, spoken interaction is fundamental to human behavior. Numerous studies have documented general discourse strategies in interpersonal, spoken communication [1,2]. Similarly, much is already known about the human factors issues associated with audiotext applications (in which telephone callers provide input via Touch-Tone to a system which responds with digitized or synthesized speech) [3,4,5]. However, there has to date been little empirical study of the manner in which a person might cooperate with a Spoken Language System (SLS). Nor is there much in the way of empirical data to suggest how an SLS can be constructed to best cooperate with the speaker.

Some Spoken Language Systems aspire to inter-human levels of interaction. Others mimic the lock-step, code-driven menu hierarchies of audiotext systems. But there is a wide spectrum of possible interaction types in between these antipodes, and the optimal balance amongst system performance, user friendliness, and financial cost and benefit associated with deploying such systems could reside anywhere along this spectrum.

Most research to date has focused on how Automatic Speech Recognition (ASR) systems will respond to speakers, in that this research looks at improving accuracy and performance under changing voices, vocabularies, transmission media, noise levels, and so on. This research, all by itself, is well-suited towards predicting the performance of single-question/single-answer applications. But as Spoken Language Systems frequently involve multiple queries and responses, we must also look at how speakers respond to Spoken Language Systems. An audiotext application driven by speech input instead of Touch-Tone will affect the caller's speech, based on the caller's perception of the system's capabilities and behavior. Will speakers grasp the distinction between the specific keywords in a system's vocabulary and their numerous semantic equivalents? How will they react if told flatly "That response was not understood," no matter what the cause of the recognition failure? Even if their speech is understood and processed, if they speak as they would to a person, but hear a prerecorded voice response from a computer, will they continue to speak "normally"?

In order to accurately control and predict an SLS's ultimate performance under real user conditions, then, both sides of the human-computer interaction must be understood: which ASR capabilities provide the most help in completing a multi-step service, and what speaking behavior can be expected from callers encountering such a system. The two questions are highly interdependent, so both must be investigated in parallel. In order to gather empirical data, we have designed a trial that models hypothetical features of SLSs, and measures users' responses to these features. This paper will address how the Money Talks trial will model various levels of speech automation technology, based on users' expectations from human-human discourse. The data collected will indicate how closely a man-machine dialogue must emulate a service provided by humans.

2. DETERMINING USER AND TECHNOLOGY REQUIREMENTS

How do users interact with speech automated systems? The answer to this question is necessary to develop and deploy appropriate speech recognition technologies to handle the target service. Knowing the range of responses to expect will ensure deployment of the appropriate level of ASR technology—or determine whether, in fact, the service is conducive to automation given currently available technology.

Laboratory investigations of the behavior of cooperative volunteers cannot adequately predict the behavior of real users in a deployable service. Paid volunteers presented with a service simulation can be polled for their perceived preferences, but measures

on their willingness to comply with instructions may be misleading. The speech outputs of volunteers in a laboratory setting also generalize poorly to speech outputs of goal-directed users accessing a real service [6].

Observations of real services offered by human representatives are often used to establish how callers typically respond. Human-human interaction, however, differs markedly from human-computer interaction. Several experiments have investigated customer responses in human-computer interaction, and have collected human-human baseline data for comparison [7,8]. Callers interacting with humans were consistently more verbose and less targeted than callers interacting with what appeared to be a computer-controlled discourse. Observing only the human-human interaction would have resulted in the erroneous conclusion that the service in question could not be automated with current ASR technology.

Studying real users interacting with automated systems will provide the best insights about user behavior and requirements. This can be accomplished through simulated automation of real services. Callers are presented with an automated interface, but an intelligent human "wizard" is at all times present, covertly directing the discourse and modeling the speech recognition technology under investigation. Observing users in simulations of automated services has revealed that users can, in fact, constrain their responses appropriately when they believe there is no intelligent human present to decipher their unsolicited additional verbiage [7,8,9].

3. HUMAN-TO-HUMAN INTERACTION

Communication through speech is frequently touted in the speech recognition literature as the most natural means of communication. The preference for talking, however, may result more from a listener's facility in interpreting speech than from the motoric differences between pressing keys and speaking. Some features of human-human discourse which our trial will simulate are:

Freedom to use semantic equivalents rather than rigidly conforming to a pre-defined "code." When using speech as the modality, semantically similar tokens are generally accepted as synonymous. Speakers/listeners accept "yes," "yeah," "uh-huh," and many other words and phrases as markers of the affirmative response, allowing the speaker/listener to focus on meaning rather than the precise signals used to transmit that meaning.

Freedom to repair speech. Speakers do not communicate using pre-rehearsed scripts. Speakers are often organizing their thoughts and translating them to speech simultaneously. As a result, natural communication is fraught with hesitations and false starts. Listeners tolerate these false starts, and speakers can repair the errors without restarting the utterance.

Limited freedom to determine discourse flow. When the speaker provides information in a "non-standard" sequence, listeners can compensate. For instance, in the Directory Assistance domain, "In New York" followed by "John Smith" would be interchangeable with "John Smith in NY." This discourse flexibility allows necessary information to be provided in whatever order the speaker deems appropriate.

Technology will not, in the near term, recognize speech incorporating these kinds of variations as well as human listeners can. SLS requirements, then, will in part be determined by the degree to which speakers can adjust and provide "compliant" input when some or all of these freedoms are absent.

4. THE MONEY TALKS TRIAL

The *Money Talks* trial will determine application requirements in a real user, service-providing environment: a telephone-based information service. Random and unsolicited callers, dialing into a current Touch-Tone driven financial information system, will be presented instead with a speech-driven interface. Each such caller will verbally traverse through a dialog defined by one of six different *scenarios*—models of machine behavior—wherein system response is constrained by type of recognition technology, reprompt strategy, and ability to categorize non-compliant input. The six scenarios will simulate: (1) isolated digit recognition; (2) embedded digit recognition; (3) isolated keyword recognition; (4) embedded keyword recognition; (5) continuous speech recognition (CSR), without discourse flexibility; and (6) CSR with discourse flexibility. Within each scenario, different callers will also hear different wordings of the system's pre-recorded prompts. When callers speak in a manner that the simulated technology can understand, they advance through the automated dialogue; when they do not, they are reprompted. As with the Touch-Tone interface, multiple "errors" on the callers' part will result in a transfer to a human representative for assistance. The caller will also be transferred to a human if they say the word "representative" at any point. The trial will digitally record all utterances, and will collect statistically relevant measures of callers' behavior. The system, then, will act as an arena within which different speech technologies and interaction styles attempt to produce the most successful interactions with callers.

5. TECHNOLOGY VARIATIONS

The Money Talks trial enjoys the advantage of a live wizard interpreting the user's input. The trial's goal, however, is to determine the technology requirements for a fully automated system that interprets callers' responses and prompts them for the next required piece of information. The wizard will in fact model the candidate speech technologies. The most accessible commercial technology recognizes digits only. Can users limit their responses to digits only, and remain satisfied with the service? Another scenario will allow pre-selected key words. Will this be more feasible and palatable for the users? Are users comfortable with digits and/or key words if they are required to speak them in isolation?

The digit and keyword scenarios assess relatively small vocabulary speech recognition systems. The fifth scenario will model a larger vocabulary system with CSR, allowing users to request information in any way they choose. Post hoc analysis of the recorded database will indicate how large a vocabulary is actually required at any point in the dialogue, and whether in fact what is required is an actual CSR capability, or simply a keyword-spotting system with a large vocabulary of synonymous content words.

Statistical measures of users' reactions will reflect to what degree callers are more amenable to such systems.

The last scenario will model a full blown language understanding system. Users of this service will not be obligated to present information in a pre-determined sequence or to use pre-selected words. Results from this scenario will answer the questions: How much variability exists in the way users present their requests when there are no obvious constraints? How large a vocabulary is required? How much more likely are users to get their required information when the system does not dictate their presentation style? Finally, how satisfied are the callers with this system?

The Money Talks trial will address many important questions about technology requirements and user satisfaction. But several questions will remain unanswered, in order to keep the trial manageable. For example, how do users respond when speech recognition technology fails? The wizards will model pre-defined types of recognition technology, but will not simulate speech recognition errors unless the calling customer has not complied with instructions. Also, barge-in capability will always be available to the customer, even when modeling the simplest technologies. (This was a stipulation from the sponsoring information service providers, who stated that the service must minimally provide the functionality offered through Touch-Tone, which does include barge-in.) As a result, customers' reactions if barge-in were not an option will not be tracked. Finally, learning effects will not be measured. The trial will tap a large and random user population, where individuals are unlikely to access the speech automated system more than once. Therefore, this trial cannot map the potentially large effect that learning and familiarity may have on customer performance and satisfaction.

6. PROMPTING VARIATIONS

In addition to modeling various forms of technology, the wording of prompting requests to callers will be varied. Several previous speech recognition trials have highlighted the important effect of prompting on customers' responses. For example, NYNEX performed a trial to assess automatability of operator handled intercept calls [7]. Various prompts were presented to customers to verbally elicit a telephone number in a way that current ASR technology could decode. Changing prompting parameters such as wordiness of the prompt or utterance speed clearly affected the callers' response mode and the likelihood that they would abandon the service. Another operator services trial was conducted to elicit isolated city names for Directory Assistance. Four prompts were compared, with respect to their ability to elicit just the requested target. The most successful of the four captured 64% of the calling population, vs. 15% for the least successful prompt [9].

When callers are not familiar with an automated service, intuition suggests that prompts should be as informative as possible, with the goal of eliciting maximal cooperation. Data from previous automated service trials belie this conclusion. In the intercept trial [7], various forms of customer greetings were presented. The greetings ranged from wordy and informative introductions to curt greetings to no greeting at all. Customers reacted in a counter-intuitive fashion: abandon rates were the highest in the case where customer cooperation was explicitly requested; the lowest abandon rates were found when the greeting was omitted. Therefore, prompts in the Money Talks trial will be as terse as possible. Lengthier prompts will be presented when callers offer invalid input, and an explanation of how to repair that input is warranted.

Just as predictions of callers' responses to an SLS might be inferred from observed human-human interactions or from patterns of audiotext use, so too might prompt wording be based on how human representatives word queries, or alternatively, on how audiotext prompts are worded. Automated prompts mimicking what representatives say would offer the most familiar interface to the caller. Such an approach, however, has proven counterproductive in previous NYNEX trials: prerecorded prompts that sounded like representatives' extemporaneous queries led to more non-target speech from callers. On the other hand, verbose mapping instructions similar to those of audiotext systems ("to continue, say 'yes'; to quit, say 'no'") do little to make a call shorter or more "natural," and may be unnecessary for more advanced ASR systems, which have less need to constrain callers' speech. They may even be unnecessary in simpler systems for queries that require closed-class answers, such as "Would you like to continue?" or "What is the dollar amount?", where a large percentage of replies may be captured with a relatively small, fixed vocabulary or grammar. In order to determine prompting requirements, then, prompt variations amongst calls in Money Talks will attempt to span this spectrum, from "naturally concise" to "prescriptively explicit."

The prompting variations in this trial address the following questions:

- Which prompts are most successful at eliciting the desired speech for each technology scenario?
- What is the most effective reprompting strategy to encourage callers to repeat their inputs, after an initial "failure"?

The answer to the first question will provide pragmatic information about prompt selection in a real service. The second question will have a more direct effect on ASR technology development. Reprompts can be generic, such as "We did not understand. Please repeat." They can also be more instructive, as in "Please speak louder." The latter example gives the user information about how to repair the input. If specific instructions result in more successful repair strategies than the generic reprompt, then it will be important for ASR technology to provide more detailed return codes than simply "recognition failure."

7. ANTICIPATED OUTPUTS FROM THE MONEY TALKS TRIAL

Data from the Money Talks trial are being collected to determine the feasibility of automating information services with speech technology, and to identify which ASR technology is most appropriate for the task. The follow up step will be to advance ASR technology accordingly. Towards this goal, users' responses will be recorded for later ASR training and testing purposes. The recorded data will also provide the exact wording of customers in response to the automated queries, allowing for more detailed follow-up analyses. This will facilitate better language modeling for this and similar applications.

Additionally, statistical data will be recorded for each call, describing such attributes as length of call, classification of response to each prompt, number of reprompts issued, manner in which the call was ended, and so on. User acceptance of the various scenarios will be measured in several ways. First, were customers compliant to particular requests, or did they opt for human intervention? When customers were willing to participate, were they able to comply with the instructions? Did they require reprompts at several junctures? Were reprompts adequate in getting users to comply? What appears to be callers' preferred mode of interaction: Were isolated digits adequate? Or do users need a spoken language understanding system to successfully and happily traverse through the service? How satisfied were users with any version of the speech driven interface, compared with a Touch-Tone managed service?

Responses to these questions will serve to motivate future efforts in ASR development to provide interactive services where the content or sequence of callers' requests is not entirely predictable. The trial results will determine whether currently available or near-term technologies are adequate to automate such services, or whether ASR development must first advance to handle features like spoken language understanding to make the service useful and desirable. Finally, the conclusions from this trial will ensure that the interface to such a system will be designed with sensitivity to users' requirements for successful human-computer interactions.

8. ACKNOWLEDGMENTS

The authors wish to acknowledge Ben Chigier, Jim Kondziela, Ed Man, John Pitrelli, and Dina Yashchin for their contributions to the design of the Money Talks project.

9. REFERENCES

[1] Dobroth, K.M., Zeigler, B.L., Karis, D. "Future Directions for Audio Interface Research: Characteristics of Human-Human Order-Entry Conversations." *AVIOS*, pp. 277-282, 1989.

[2] Green, G.M. *Pragmatics and Natural Language Understanding*. Hillsdale: Lawrence Erlbaum Associates.1989.

[3] Gould, J.D. and Boies, S.L. "Human Factors Challenges in Creating a Principal Support Office System: The Speech Filing System Approach." *ACM Transactions on Office Information Systems*, 1(4), pp. 273-298, 1983.

[4] Information Industry Association, *Voice Messaging User Interface Specification*, Washington, D.C. 1990.

[5] Pollard, D. and Cooper, M.B. "The Effect of Feedback on Keying Performance." *Applied Ergonomics*, 10(4), pp. 194-200, 1979.

[6] Spitz, J. and the AI Speech Technology Group. "Collection and Analysis of Data from Real Users: Implications for Speech Recognition/Understanding Systems." *DARPA Speech and Natural Language Workshop*, pp. 164-169, 1991.

[7] Yashchin, D., Basson, S., Lauritzen, N., Levas, S., Loring, A., Rubin-Spitz, J. "Performance of Speech Recognition Devices: Evaluating Speech Produced Over the Telephone Network." *ICASSP*, pp. 552-555, 1989.

[8] Basson, S., Christie, O., Levas, S., Spitz, J. "Evaluating Speech Recognition Potential in Automating Directory Assistance Call Completion." *AVIOS*, 1989.

[9] Chigier, B., and Spitz, J. "Are Laboratory Databases Appropriate for Training and Testing Telephone Speech Recognizers?" *ICSLP*, pp. 1017-1020, 1990.