A SIMULATION-BASED RESEARCH STRATEGY FOR DESIGNING COMPLEX NL SYSTEMS*

Sharon Oviatt, Philip Cohen, Michelle Wang & Jeremy Gaston†

Computer Dialogue Laboratory
A.I. Center, SRI International
333 Ravenswood Avenue
Menlo Park, California, U.S.A. 94025

ABSTRACT

Basic research is critically needed to guide the development of a new generation of multimodal and multilingual NL systems. This paper summarizes the goals, capabilities, computing environment, and performance characteristics of a new semi-automatic simulation technique. This technique has been designed to support a wide spectrum of empirical studies on highly interactive speech, writing, and multimodal systems incorporating pen and voice. Initial studies using this technique have provided information on people's language, performance, and preferential use of these communication modalities, either alone or in multimodal combination. One aim of this research has been to explore how the selection of input modality and presentation format can be used to reduce difficult sources of linguistic variability in people's speech and writing, such that more robust system processing results. The development of interface techniques for channeling users' language will be important to the ability of complex NL systems to function successfully in actual field use, as well as to the overall commercialization of this technology. Future extensions of the present simulation research also are discussed.

1. INTRODUCTION

Basic research is critically needed to guide the development of a new generation of complex natural language systems that are still in the planning stages, such as ones that support multimodal, multilingual, or multiparty exchanges across a variety of intended applications. In the case of planned multimodal systems, for example, the potential exists to support more robust, productive, and flexible human-computer interaction than that afforded by current unimodal ones [3]. However, since multimodal systems are relatively complex, the problem of how to design optimal configurations is unlikely to be solved through simple intuition alone. Advance empirical work with human subjects will be needed to generate a factual basis for designing multimodal systems that can actually deliver performance superior to unimodal ones.

In particular, there is a special need for both methodological tools and research results based on high-quality simulations of proposed complex NL systems. Such simulations can reveal specific information about people's language, task performance, and preferential use of different types of systems, so that they can be designed to handle expected input. Likewise, simulation research provides a relatively affordable and nimble way to compare the specific advantages and disadvantages of alternative architectures, such that more strategic designs can be developed in support of particular applications. In the longer term, conclusions based on a series of related simulation studies also can provide a broader and more principled perspective on the best application prospects for emerging technologies such as speech, pen, and multimodal systems incoporating them.

In part for these reasons, simulation studies of spoken language systems have become common in the past few years, and have begun to contribute to our understanding of human speech to computers [1, 5, 6, 7, 8, 17]. However, spoken language simulations typically have been slow and cumbersome. There is concern that delayed responding may systematically distort the data that these simulation studies were designed to collect, especially for a modality like speech from which people expect speed [6, 10, 15]. Unlike research on spoken language systems, there currently is very little literature on handwriting and pen systems. In particular, no simulation studies have been reported on: (1) interactive handwriting[1] [6], (2) comparing interactive speech versus handwriting as alternative ways to interact with a system, or (3) examining the combined use of speech and handwriting to simulated multimodal systems of different types. Potential advantages of a combined pen/voice system have been outlined previously [4, 12]. High quality simulation

*This research was supported in part by Grant No. IRI-9213472 from the National Science Foundation to the first authors, as well as additional funding and equipment donations from ATR International, Apple Computer, USWest, and Wacom Inc. Any opinions, findings, or conclusions expressed in this paper are those of the authors, and do not necessarily reflect the views of our sponsors.

†Michelle Wang is affiliated with the Computer Science Department and Jeremy Gaston with the Symbolic Systems Program at Stanford University.

[1] Although we are familiar with noninteractive writing from everyday activities like personal notetaking, very little is known about *interactive* writing and pen use as a modality of human-computer interaction.

research on these topics will be especially important to the successful design of mobile computing technology, much of which will emphasize communications and be keyboardless.

The simulation technique developed for this research aims to: (1) support a very rapid exchange with simulated speech, pen, and pen/voice systems, such that response delays are less than 1 second and interactions can be subject-paced, (2) provide a tool for investigating interactive handwriting and other pen functionality, and (3) devise a technique appropriate for comparing people's use of speech and writing, such that differences between these communication modalities and their related technologies can be better understood. Toward these ends, an adaptable simulation method was designed that supports a wide range of studies investigating how people speak, write, or use both pen and voice when interacting with a system to complete qualitatively different tasks (e.g., verbal/temporal, computational/numeric, graphic/cartographic). The method also supports examination of different issues in spoken, written, and combined pen/voice interactions (e.g., typical error patterns and resolution strategies).

In developing this simulation, an emphasis was placed on providing automated support for streamlining the simulation to the extent needed to create facile, subject-paced interactions with clear feedback, and to have comparable specifications for the different modalities. Response speed was achieved in part by using scenarios with correct solutions, and by preloading information. This enabled the assistant to click on predefined fields in order to respond quickly. In addition, the simulated system was based on a conversational model that provides analogues of human backchannel and propositional confirmations. Initial tasks involving service transactions embedded propositional-level confirmations in a compact transaction "receipt," an approach that contributed to the simulation's clarity and speed. Finally, emphasis was placed on automating features to reduce attentional demand on the simulation assistant, which also contributed to the fast pace and low rate of technical errors in the present simulation.

2. SIMULATION METHOD

Basic simulation features for the studies completed to date are summarized below, and have been detailed elsewhere [16], although some adaptations to these specifications are in progress to accommodate planned research.

2.1. Procedure and Instructions

Volunteer participants coming into the Computer Dialogue Laboratory at SRI are told that the research project aims to develop and test a new pen/voice system for use on future portable devices. To date, subjects have included a broad spectrum of white-collar professionals, excluding computer scientists. All participants so far have believed that the "system" was a fully functional one. Following each session, they are debriefed about the nature and rationale for conducting a simulation.

During the study, subjects receive written instructions about how to enter information on an LCD tablet when writing, when speaking, and when free to use both modalities. When writing, they are told to handwrite information with the electronic stylus directly onto active areas on the tablet. They are free to print or write cursive. When speaking, subjects are instructed to tap and hold the stylus on active areas as they speak into the microphone. During free choice, people are completely free to use either modality in any way they wish. Participants also receive written instructions about how to use the system to complete realistic tasks, which currently focus on the broad class of service-oriented transactions (e.g., car rental reservations, personal banking, real estate selection). Then they practice several scenarios using spoken and written input until the system and the tasks are completely clear.

People are encouraged to speak and write naturally. They are asked to complete the tasks according to instructions, while working at their own pace. Other than providing motivation to complete the tasks and specifying the input modality, an effort is made not to influence the specific manner in which subjects express themselves. They are encouraged to focus on completing the tasks and are told that, if their input cannot be processed for any reason, this will be clear immediately since the system will respond with ??? to prompt them to try again. Subjects are told how to remove or replace information as needed. Otherwise, they are told that input will be confirmed by the system on a transaction receipt, which they can monitor to check that their requests are being met (see next section for details). Of course, participants' input actually is received by an informed assistant, who performs the role of interpreting and responding as the system would.

The simulation assistant is instructed to respond as accurately and rapidly as possible to any spoken or written information corresponding to predefined receipt fields. Essentially, the assistant tracks the subject's input, clicking with a mouse on predefined fields on a Sun SPARCstation to send confirmations back to the subject. Under some circumstances, the assistant is instructed to send a ??? prompt instead of a confirmation. For example, subjects receive ??? feedback when input is judged to be inaudible or illegible, when the subject forgets to

supply task-critical information, or when input clearly is inappropriate, ambiguous, or underspecified. In general, however, the assistant is instructed to use ??? feedback sparingly in order to minimize intervention with people's natural tendencies to speak or write. If the subject commits a procedural error, such as forgetting to click before entering speech or attempting to enter information using the wrong modality, then the assistant is instructed not to respond until the subject recovers and correctly engages the system. The assistant's task is sufficiently automated that he or she is free to focus attention on monitoring the accuracy of incoming information, and on maintaining sufficient vigilance to respond promptly with confirmations.

2.2. Presentation Format

For studies completed to date, two different prompting techniques have been used to guide subjects' spoken and written input— one unconstrained and one forms-based. In the relatively unconstrained presentation format, subjects must take the initiative to ask questions or state needs in one general workspace area. No specific system prompts direct their input. They simply continue providing information until their transaction receipt is completed, correctly reflecting their requests. In this case, guidance is provided primarily by the task itself and the receipt. When the presentation format is a form, labeled fields are used to elicit specific task content, for example: **Car pickup location**[ ]. In this case, the interaction is more system-guided, and linguistic and layout cues are used to channel the content and order of people's language as they work.

For other studies in which people work with visual information (e.g., graphic/cartographic tasks), different graphic dimensions of presentation format are manipulated. In all studies, the goal is to examine the impact of presentation format on people's language and performance as they either speak or write to a simulated system. As a more specific aim, assessments are being conducted of the extent to which different formats naturally constrain linguistic variability, resulting in opportunities for more robust natural language processing.

2.3. Conversational Feedback

With respect to system feedback, a conversational model of human-computer interaction was adopted. As a result, analogues are provided of human backchannel and propositional-level confirmations. These confirmations function the same for different input modalities and presentation formats. With respect to backchannel signals, subjects receive *** immediately following spoken input, and an electronic ink trace following written input. These confirmations are presented in the tablet's active area or a narrow "confirmation panel" just below it. Subjects are told that this feedback indicates that their input has been legible/audible and processable by the system, and that they should continue.

In addition to this backchannel-level signal, subjects are told to verify that their requests are being met successfully by checking the content of the receipt at the bottom of the tablet. This receipt is designed to confirm all task-critical information supplied during the interaction, thereby providing propositional confirmations. It remains visible throughout the transaction, and is completed gradually as the interaction proceeds. Although the receipt varies for different tasks, its form and content remains the same for different modalities and presentation formats.

Apart from confirmation feedback, the simulation also responds to people's questions and commands by transmitting textual and tabular feedback. For example, if a subject selects the car model that he or she wants and then says, "Do you have infant seats?" or "Show me the car options," a brief table would be displayed in which available items like infant seats and car phones are listed along with their cost.

2.4. Automated Features

To simplify and speed up system responding, the correct receipt information associated with each task is preloaded for the set of tasks that a subject is to receive. A series of preprogrammed dependency relations between specified task-critical information and associated receipt fields is used to support the automation of propositional confirmations. As mentioned earlier, with this arrangement the assistant simply needs to click on certain predefined fields to send appropriate acknowledgments automatically as the subject gradually supplies relevant information. Of course, if the subject makes a performance error, the assistant must manually type and confirm the error that occurs. In such cases, however, canonical answers are maintained so that they can be confirmed quickly when people self-correct, which they tend to do over 50% of the time. The automated simulation strategy described above works well when research can take advantage of task scenarios that entail a limited set of correct answers.

An additional automated feature of the present simulation technique is a "random error generator," which is designed to ensure that subjects encounter at least a minimal level of simulated system errors, in part to support the credibility of the simulation. In this research, if subjects do not receive at least one ??? response from

the system during a set of two tasks, then the simulation generates one. This results in a minimum baseline rate of one simulated error per 33 items of information supplied, or 3%, which in this research has been considered a relatively error-free environment. The simulated errors are distributed randomly across all task-critical information supplied for the set of tasks.

2.5. Performance Characteristics

The described method for organizing simulated response feedback was responsible in part for the fast pace of the present simulation. In studies conducted to date, response delays during the simulation have averaged 0.4 second between a subject's input and visible confirmation on the tablet receipt, with less than a 1-second delay in all conditions. The rate of technical errors in executing the assistant's role according to instructions has been low, averaging 0.05 such errors per task. Furthermore, any major error by the assistant would result in discarding that subject's data, which currently has been averaging 6% of subjects tested. The present simulation also appears to be adequately credible, since no participants to date have doubted that it was a fully functional system. As a result, no data has been discarded for this reason.

2.6. Simulation Environment

The computing equipment that supports this simulation technique includes two Sun workstations, one a SPARCstation 2, that are linked via ethernet. A Wacom HD-648A integral transparent digitizing tablet/LCD display is interfaced to the SPARC 2 through a Vigra S-bus VGA card. An accompanying cordless digitizing pen is used for writing, clicking to speak, pointing, or otherwise operating the tablet. A Crown PCC 160 microphone transmits spoken input from the subject to the simulation assistant, who listens through a pair of stereo speakers from a remote location. The assistant also views an image of the subject working at the tablet, along with an image of all visible input and feedback occurring on the tablet.

The user interface is based on the X-windows system, employing MIT Athena widgets. X-windows is used for its ability to display results on multiple screens, including the subject's tablet and the assistant's workstation, and because the resulting program runs on equipment from several manufacturers. Two aspects of the system architecture are designed for rapid interface adaptability. First, Widget Creation Language (WCL) enables non-programmers to alter the user interface layout. Second, a simple textual language and interpreter were created to enable declarative specification of widget behavior and interrelations. Some widget behavior also is written in the C programming language.

Various modifications to the standard X-windows operation have been deployed to ensure adequate real-time responding needed for acceptable handwriting quality and speed. To avoid objectionable lag in the system's electronic ink echo, a high-performance workstation (i.e., Sun SPARCstation 2) is used to process the subject's input.

2.7. Data Capture

With respect to data collection, all human-computer interactions are videotaped for subsequent analysis. The recording is a side-by-side split-screen image, created using a JVC KM-1200U special-effects generator. Videotaping is conducted unobtrusively with a remote genlocked Panasonic WV-D5000 videocamera filming through a one-way mirror. Data capture includes a close-up of the subject working at the LCD tablet, and a real-time record of interactions on the tablet, including the subject's input, simulated feedback, and the gradually completed receipt. This image is recorded internally from the assistant's workstation, is processed through a Lyon Lamb scan converter, and then is merged using the special-effects generator and preserved on videotape for later analysis. In addition to being transmitted to the simulation assistant, the subject's speech is recorded and stored in analog form on a timecoded videotape, and later is transcribed for data analysis. All handwritten input is recorded on-line during real-time tablet interactions, which then is preserved on videotape and available for hardcopy printout.

3. RESEARCH DESIGN

In studies conducted at SRI to date, the experimental design usually has been a completely-crossed factorial with repeated measures, or a within-subjects design. Primary factors of interest have included: (1) communication modality (speech-only, pen-only, combined pen/voice), and (2) presentation format (form-based, unconstrained). In a typical study, each subject completes a series of 12 tasks, two representing each of the six main conditions. The order of presenting conditions is counterbalanced across subjects.

This general design has been selected for its relative efficiency and power and, in particular, for its ability to control linguistic variability due to individual differences. In brief, for example, this design permits comparing how the *same* person completing the *same* tasks displays one type of language and performance while speaking, but then switches this language and performance when writing.

4. SAMPLE RESULTS

The variability inherent in people's language, whether spoken or written, poses a substantial challenge to the successful design of future NL systems. One aspect of this research has been a comprehensive assessment of the linguistic variability evident in people's speech and writing at various levels of processing, including acoustic, lexical, syntactic, and semantic. Full reports of these results are forthcoming [11, 14]. Special emphasis has been placed on identifying problematic sources of variability for system processing, as well as an explanation of the circumstances and apparent reasons for their occurrence. In connection with these analyses, one goal of this research program has been to identify specific interface techniques that may naturally channel users' language in ways that reduce or eliminate difficult sources of variability, so that more robust system processing can be achieved. In particular, the impact of selecting a particular input modality or presentation format is being examined, so that future system designers will have the option of choosing a particular modality or format because doing so will minimize expected performance failures of their planned NL systems.

To briefly illustrate the research theme of reducing linguistic variability through selection of modality and format, the results of an analysis related to syntactic ambiguity are summarized. Two indices of relative ambiguity were measured for all phrasal and sentential utterances that people spoke to an unconstrained format (SNF), wrote in an unconstrained format (WNF), spoke to a form (SF), or wrote in a form (WF). Two different estimates of parse ambiguity were computed to check for convergence of results. First, utterances produced under the different simulation conditions were parsed using DIALOGIC [9], a robust text processing system developed at SRI that employs a broad coverage grammar. Second, a summary was computed of the number of *canonical* parses produced by DIALOGIC, through a mapping of each DIALOGIC parse to an emerging national standard parse tree representation called PARSEVAL form[2] [2]. The average number of DIALOGIC and PARSEVAL parses generated per utterance for the different simulation conditions is summarized in Table 1, along with the percentage of all utterances in each condition that were phrases or sentences and therefore appropriate for parsing.

None of the subjects produced phrases or sentences when writing to a form, so none of the simple utterances from this condition were appropriate for parsing. The percentage of phrase and sentential utterances available for parsing was greater for unconstrained than form-based input, and greater for spoken than written input. Comparison of both parse metrics for unconstrained and form-based speech revealed that using a form significantly reduced the average number of parses per utterance, t (paired) = 2.50 (df = 5), p < .03, one-tailed (DIALOGIC), and t (paired) = 2.35 (df = 5), p < .04, one-tailed (PARSEVAL). When comparisons were made of the same subjects accomplishing the same tasks, the parse ambiguity of utterances in the unconstrained format averaged 232% higher for DIALOGIC and 157% higher for PARSEVAL than when communicating to a form. However, comparison of both parse metrics for speech and writing in an unconstrained format did not confirm that use of the written modality reduced the average number of parses per utterance, t (paired) = 1.18 (df = 14), p > .10, one-tailed (DIALOGIC), and t < 1 (PARSEVAL). That is, reliable reduction of parse ambiguity was obtained only through manipulation of the presentation format.

COND.	DIALOGIC	PARSEVAL	UTTERANCES PARSED
SNF	20.9	7.2	36%
WNF	10.7	4.4	18%
SF	6.3	2.8	8%
WF	—	—	0%

Table 1: Average number of DIALOGIC and PARSEVAL parses per utterance as a function of modality and format.

This pattern of results suggests that selection of presentation format can have a substantial impact on the ease of natural language processing, with direct implications for improved system robustness. In addition, post-experimental interviews indicated that participants preferred form-based interactions over unconstrained ones by a factor of 2-to-1 in the present tasks. In particular, both the guidance and assurance of completeness associated with a form were considered desirable. This indicates that the *a priori* assumption that any type of constraint will be viewed by people as unacceptable or unnatural clearly is not always valid. Furthermore, such a presumption may simply bias system development away from good prospects for shorter-term gain. The application of this kind of interface knowledge will be important to the successful performance and commercialization of future natural language technology.

[2]PARSEVAL form is designed to reflect agreement among computational linguists simply on the major constituent bracketings, so PARSEVAL identification of syntactic structures should tend to represent the commonalities among many different systems.

5. FUTURE DIRECTIONS

The long-term goal of the present research method is to support a wide spectrum of advance empirical studies on interactive speech, pen, and pen/voice systems under different circumstances of theoretical and commercial interest. Future extensions of the present simulation research are under way to examine issues relevant to multilingual and other multiparty applications [13]. In addition, a taxonomy of tasks is being developed in order to establish a more analytical basis for distinguishing when findings do or do not generalize to qualitatively different domains, such that future work need not approach each new application as an unknown entity. Efforts also are under way to define the important dimensions of system interactivity, such as feedback characteristics and error resolution strategies, as well as their impact on human-computer interaction. Finally, in addition to providing proactive guidance for system design, a further aim of this simulation research is to yield better information about the range of preferred metrics for conducting performance assessments of future NL systems, including their accuracy, efficiency, learnability, flexibility, ease of use, expressive power, and breadth of utility.

6. ACKNOWLEDGMENTS

Thanks to John Dowding, Dan Wilk, Martin Fong, and Michael Frank for invaluable programming assistance during the design and adaptation of the simulation. Special thanks also to Dan Wilk and Martin Fong for acting as the simulation assistant during experimental studies, to Zak Zaidman for general experimental assistance, and to John Bear, Jerry Hobbs, and Mabry Tyson for assisting with the preparation of DIALOGIC and PARSEVAL parses. Finally, thanks to the many volunteers who so generously offered their time to participate in this research.

References

1. F. Andry, E. Bilange, F. Charpentier, K. Choukri, M. Ponamalé, and S. Soudoplatoff. Computerised simulation tools for the design of an oral dialogue system. In *Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218).* Commission of the European Communities, 1990.
2. E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop,* pages 306–311. Morgan Kaufmann, Inc., February 1991.
3. R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Bierman, M. Bush, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorain, and V. Zue. NSF workshop on spoken language understanding. Technical Report CS/E 92-014, Oregon Graduate Institute, September 1992.
4. H. D. Crane. Writing and talking to computers. Business Intelligence Program Report D91-1557, SRI International, Menlo Park, California, July 1991.
5. N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz studies — why and how. In L. Ahrenberg, N. Dahlbäck, and A. Jönsson, editors, *Proceedings from the Workshop on Empirical Models and Methodology for Natural Language Dialogue Systems,* Trento, Italy, April 1992. Association for Computational Linguistics, Third Conference on Applied Natural Language Processing.
6. N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech and Language,* 5(1):81–99, 1991.
7. M. Guyomard and J. Siroux. Experimentation in the specification of an oral dialogue. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems.* Springer Verlag, Berlin, B. R. D., 1988. NATO ASI Series, vol. 46.
8. C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The ATIS spoken language systems pilot corpus. In *Proceedings of the 3rd Darpa Workshop on Speech and Natural Language,* pages 96–101, San Mateo, California, 1990. Morgan Kaufmann Publishers, Inc.
9. J. R. Hobbs, D. E. Appelt, J. Bear, M. Tyson, and D. Magerman. Robust processing of real-world natural-language texts. In P. S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval.* Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey, 1992.
10. A. F. Newell, J. L. Arnott, K. Carter, and G. Cruickshank. Listening typewriter simulation studies. *International Journal of Man-machine Studies,* 33(1):1–19, 1990.
11. S. L. Oviatt. Writing and talking to future interactive systems. manuscript in preparation.
12. S. L. Oviatt. Pen/voice: Complementary multimodal communication. In *Proceedings of Speech Tech'92,* pages 238–241, New York, February 1992.
13. S. L. Oviatt. Toward multimodal support for interpreted telephone dialogues. In M. M. Taylor, F. Néel, and D. G. Bouwhuis, editors, *Structure of Multimodal Dialogue.* Elsevier Science Publishers B. V., Amsterdam, Netherlands, in press.
14. S. L. Oviatt and P. R. Cohen. Interface techniques for enhancing robust performance of speech and handwriting systems. manuscript in preparation.
15. S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language,* 5(4):297–326, 1991a.
16. S. L. Oviatt, P. R. Cohen, M. W. Fong, and M. P. Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *Proceedings of the 1992 International Conference on Spoken Language Processing,* Banff, Canada, October 1992.
17. E. Zoltan-Ford. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies,* 34:527–547, 1991.