

INTEGRATION THEMES IN MULTIMODAL HUMAN-COMPUTER INTERACTION*

Sharon Oviatt and Erik Olsen

Computer Dialogue Laboratory & Artificial Intelligence Center
SRI International, 333 Ravenswood Avenue, Menlo Park, CA., U. S. A. 94025

ABSTRACT

This research examines how people integrate spoken and written input during multimodal human-computer interaction. Three studies used a semi-automatic simulation technique to collect data on people's free use of spoken and written input. Within-subject repeated-measures studies were designed, with data analyzed from 44 subjects and 240 tasks. The primary factors that govern people's selection to write versus speak at given points during a human-computer exchange were evaluated. Analyses revealed that people write digits more often than textual content, and proper names more often than other text. A form-based presentation, in comparison with an unconstrained format, also increased the likelihood of writing. However, the most influential factor in patterning people's integrated use of speech and writing is *contrastive functionality*, or the use of spoken and written input in a contrastive way to designate a shift in content or functionality, such as original versus corrected input, data versus command, and digits versus text. Different patterns of contrastive mode use accounted for approximately 57% of the integrated pen/voice use observed in these studies. Information also is summarized on preferential mode use, and simultaneity of pen/voice input. One long-term goal of this research is the development of quantitative predictive models of natural modality integration, which could provide guidance on the strategic design of robust multimodal systems.

INTRODUCTION

Since multimodal systems are relatively complex, designing them to successfully deliver natural, productive, and robust performance is unlikely to occur through simple intuition alone. In the case of multimodal systems based on human language technology, it will be particularly advantageous for system design to leverage from people's existing language patterns, since many language skills are automatized, entrenched, and not under full conscious control (e.g., intonation, timing, disfluencies). That is, to the extent that people are not able to adapt language delivery

*This research was supported by Grant No. IRI-9213472 from the National Science Foundation, contracts from AT&T/NCR, ETRI, and ATR International to SRI International, and equipment donations from Apple Computer, Sun Microsystems, and Wacom Inc. **First author's current address:** Department of Computer Science, Oregon Graduate Institute of Science and Technology, 19600 N. W. Von Neumann Drive, Beaverton, Oregon 97006, U. S. A.

fully and system design conflicts with natural usage, then the potential for such mismatches to trigger system failure increases. To design multimodal systems with performance advantages over unimodal ones, research will be needed on how people select and integrate different modalities in the context of different types of human-computer interaction. Although there is considerable current interest in developing multimodal and multimedia systems [1, 2], proactive empirical research aimed at designing well-integrated systems capable of supporting rather than fragmenting user behavior has been lacking.

The goal of the present research is to begin specifying how people integrate their use of spoken and written input during multimodal human-computer exchanges. In particular, the present simulation studies aim to identify the primary factors that influence people's selection to write or speak at given points during such an exchange — including type of content, length of content, presentation format of the interaction, and so forth. These studies also explore the overall patterning and causal basis of integrated modality use. Finally, simultaneity of pen/voice input is examined, as are users' self-reported preferences to use multimodal versus unimodal input. The results of three studies are outlined in which it was possible to analyze data in qualitatively different content domains.

METHOD

Subjects, Tasks, and Procedure

Forty-four subjects participated in this research as paid volunteers. Participants represented a broad spectrum of white-collar professionals, excluding computer scientists, and all were native speakers of English. A "Service Transaction System" was simulated that could assist users with verbal/temporal tasks (e.g., conference registration, car rental exchanges) and computational/numeric tasks (e.g., personal banking, scientific calculations).

During this research, subjects first received a general orientation to the Service Transaction System, and then were given practice using it to complete tasks. They received instructions on how to enter information on an LCD tablet when writing, speaking, and free to use both modalities. When writing, they were free to use cursive handwriting or printing, and were told to write information with the cordless electronic stylus directly onto highlighted areas on the

LCD tablet. When speaking, subjects were instructed to tap and hold the stylus on active areas as they spoke into the microphone. During free choice, they were completely free to use either modality in any way they wished. In all cases, they were encouraged to speak and write naturally, and to work at their own pace.

People also were instructed on completing tasks in two different presentation formats. In an unconstrained format, they had to take the initiative to ask questions or state needs, which they could express in an open workspace. No specific system prompts were used to direct their spoken or written input. People simply continued providing information while the system responded interactively with confirmations. During other interactions, the presentation format was explicitly structured, with linguistic and graphical cues used to structure the content and order of people's input as they worked. More specifically, labeled fields were used to elicit information (e.g., **Car pickup location**). In both cases, people continued providing information until a transaction receipt was completed at the bottom of their tablet, correctly reflecting their requests.

Other than specifying the input modality and format, an effort was made not to influence the manner in which people expressed themselves. Subjects' input actually was received by an informed assistant, who performed the role of interpreting and responding as a fully functional system would. Essentially, the assistant tracked the subject's written or spoken input, and clicked on predefined fields at a Sun SPARCstation to send confirmations back to the subject.

Semiautomatic Simulation Method

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 input modalities. In the present simulation environment, response delays averaged 0.4 second, with less than a 1-second delay in all conditions. In addition, the simulation was organized to transmit analogues of human backchannel and propositional confirmations, with propositional-level confirmations embedded in a compact transaction receipt. The simulation also was designed to be sufficiently automated so that the assistant could concentrate attention on monitoring the accuracy of incoming information, and on maintaining sufficient vigilance to ensure prompt responding. This semi-automation contributed to the fast pace of the simulation, and to a low rate of technical errors. Details of the simulation technique and its capabilities have been presented elsewhere [3, 4].

Research Design and Data Capture

Three studies were completed in which the research design was a completely crossed factorial with repeated measures. In two studies, the main factors of interest included: (1) communication modality – speech-only, pen-only, combined pen/voice, and (2) presentation format –

structured, unconstrained. One of these studies focused on verbal/temporal content, and the other on computational/numeric exchanges. In a third study on the verbal/temporal content, the communication modality was combined pen/voice throughout, but the presentation format alternated between highly structured and unconstrained.

For present purposes, only data during combined pen/voice input were analyzed, which totaled 240 tasks. All human-computer interactions were videotaped. Hard-copy transcripts also were created, with the subject's handwritten input captured automatically, spoken input transcribed verbatim onto the printouts, and all input sequenced and annotated as needed for analysis purposes.

Transcript Coding

Coding was conducted for the dependent measures listed below in the combined pen/voice conditions.

Ratio of Written Input to Total The percentage of all written words out of the total of written plus spoken words was calculated during the combined pen/voice condition for: (1) digits versus non-digit textual material, (2) proper names versus other non-digit text, (3) form-based versus unconstrained interactions, (4) digit-based content graduated in length (i.e., including 1, 3, 5, 10, and 13-digit-string input), and textual content classified as short (i.e., average letter length of 8) versus lengthy (i.e., average 21 letters), and (5) digits ranked-ordered according to perceived importance during a post-experimental interview (e.g., transaction date least important, transaction amount most important).

Contrastive Functional Use of Modes To test the hypothesis that people might organize their use of speech and writing to indicate shifts in contrastive functionality, several holistic patterns of pen/voice use were coded whenever spoken and written input both occurred. For example, during individual computations (e.g., addition, multiplication) or banking transactions (paying bills, transferring funds) in the computational/numeric study, the following dependent measures were coded: (1) original input/correction – when a computation or transaction involved both original and corrected input, the pattern of using one modality for all original input while reserving the other for corrected input, (2) data/command – for computations, the pattern of using one modality for digits and computational signs while reserving the other for issuing a command to request the total, and (3) digit/text – for banking transactions, the pattern of presenting digits and signs in one mode (e.g., transaction number, credit/debit sign) while reserving the other mode to communicate textual content (e.g., transaction recipient field).

Simultaneous Mode Use The number of times that individual subjects simultaneously spoke and wrote information was totaled for different kinds of content during each simulation. These data then were converted to a percentage of the total words.

Preference for Modalities During post-experimental interviews, people’s self-reported preference was assessed for communicating in either combined pen/voice, unimodal speech, or unimodal writing.

RESULTS

Ratio of Written to Total Input

Overall, people wrote 13% and spoke 87% of all words conveyed during the verbal/temporal simulations. The percentage of writing increased to 18% when computational/numeric content was communicated, with 82% spoken. In general, the majority of information was spoken, but people selectively interspersed written input at certain points during the interaction.

Whereas 9.7% of all textual content was written during the verbal/temporal exchanges, the percentage of writing increased to 14.7% when conveying digits. A Wilcoxon Signed Ranks test revealed that people were significantly more likely to render digits in written form than text, $T+ = 59$ ($N = 11$), $p < .02$, two-tailed. Analyses of the percentage of written digits versus text during the computational/numeric tasks replicated this finding, $T+ = 85$ ($N = 14$), $p < .045$, two-tailed. The highest rate of written to total input, or 28%, was observed during scientific calculation subtasks, which were mathematical-visual tasks comprised almost exclusively of digits and symbols.

Likewise, whereas only 6.9% of all content in the unconstrained interactions was written, the percentage of writing increased to 18.9% for structured form-based interactions. Figure 1 illustrates the percentage of written words of the total for digits versus text during both form-based and unconstrained presentation formats for the verbal/temporal simulations. A Wilcoxon Signed Ranks test revealed that people wrote more when interacting with a form than when left unconstrained, $T+ = 77.5$ ($N = 13$), $p < .03$, two-tailed.

In comparison with 9.7% written input overall for textual content, people chose to write proper names 21.5% of the time, which also was a significant elevation, Wilcoxon Signed Ranks test, $T+ = 26$ ($N = 7$), $p < .05$, two-tailed. The presence of 40% foreign surnames, which are common in the United States but potentially more difficult to pronounce, may have contributed to this elevation in written surnames.

No effect was found, however, of people selectively writing content that was brief. The percentage of written digits of graduated length, varying from brief to lengthy (i.e., 1 to 13 digits) remained between 19% and 17%, with no significant difference. Likewise, the percentage of written textual input graduated from 8 to 21 letters was 14% to 13%, respectively, which again was not significantly different. That is, although writing took considerably longer than speaking, and although people frequently abbreviated written input with standard and nonstandard abbreviations [5], they nonetheless did not selectively write briefer content. Finally, perceived importance of content did not significantly influence the likelihood that people

would choose to write something.

Contrastive Functionality

For all analyses of contrastive functional use of modes, a baseline probability for the expected pattern was calculated for each subject and compared with observed cases of the predicted pattern for that subject. For each of the three types of contrastive multimodal patterning that were examined, a Wilcoxon Signed Ranks test then was conducted to evaluate whether the predicted contrastive pattern exceeded the baseline significantly.

For original versus corrected input, people who corrected their own or a simulated error while using combined pen/voice input were significantly more likely than chance to use modalities contrastively to distinguish original input from corrections, $T+ = 39$ ($N = 9$), $p < .03$, one-tailed. People who combined spoken and written input during a banking transaction also were significantly more likely than chance to use modes contrastively to distinguish digits from text, $T+ = 45$ ($N = 10$), $p < .05$, one-tailed.

Figure 2 illustrates a contrastive pattern of modality use during a scientific calculation, in which the subject writes data (i.e., digits and computational sign), but shifts to speaking the command “go” as a request for the total. People who exercised speech and writing during a computation also were significantly more likely than chance to use modalities contrastively to distinguish data from commands, $T+ = 63$ ($N = 11$), $p < .003$, one-tailed.

These three types of contrastive patterning did not involve rigid linking of a specific mode with a specific role. Rather, some degree of symmetry was found in each case. For example, written input and spoken correction versus

was written, with the subject simultaneously subvocalizing part of the digit. In other cases, the entire digit was clearly articulated and written using both modes, often after a simulated system error, pen erasure, and so forth.

Self-Reported Modality Preference

When communicating verbal/temporal task content, 56% of people preferred to use combined pen/voice input, as opposed to unimodal speech or writing. However, when the content being communicated was computational/numeric, 89% of people preferred using multimodal pen/voice rather than unimodal input.

CONCLUSIONS

The present research evaluated natural modality integration during simulated human-computer interaction involving spoken and written input. Among the factors determining people's selection to write versus speak at given points during an exchange were:

- Task content — digits had a higher likelihood of being written than text, and proper names were more often written than other textual content.
- Presentation format — form-based interactions contained a higher percentage of written input, in comparison with unconstrained formats.
- Contrastive functionality — 57% of all integrated patterns of pen/voice use could be accounted for by people's predilection to shift modes as an indication of changing content or communicative function, such as a change between: (1) original input and correction, (2) data and command, (3) digits and text, and (4) digits and referring descriptions.

In addition, simultaneous pen/voice input was rare, occurring on fewer than 1% of all words, but typically involving digits when it did occur. Multimodal pen/voice input was preferred over either unimodal speech or writing, especially for tasks involving digits.

With respect to future directions, modality use and integration issues currently are being investigated during human-computer interaction involving complex visual displays, such as maps and photographs. One long-term goal of this research is the development of quantitative predictive models of natural modality integration, which can provide guidance on the strategic design of robust multimodal systems.

ACKNOWLEDGMENTS

Sincere thanks to the generous people who volunteered to participate in this research as subjects. Thanks also to Michael Frank, Martin Fong, and John Dowding for programming the simulation environment, to Martin Fong and Dan Wilk for playing the role of the simulation assistant during testing, and to Jeremy Gaston, Zak Zaidman, and Aaron Hallmark for careful preparation of transcripts.

spoken input and written correction each had a 50% likelihood of occurrence. Written data and spoken command had a 73% likelihood, versus 27% for spoken data and written command. Spoken text and written digits had an 85% likelihood of occurrence, versus 15% for written text and spoken digits.

One additional type of contrastive functionality, which occurred too infrequently to permit statistical analysis, involved writing digits not displayed on the screen, versus speaking descriptions referring to in-view digits (e.g., written "124.59" versus spoken "Account A balance"). In computations involving both digits and referring descriptions where spoken and written input were observed, speech always was used to convey referring expressions, whereas digits were written. That is, for this particular contrastive pattern, each modality was linked to a specific functional use.

Analysis of all computations and transactions in which speech and writing both occurred revealed that 57% displayed one of the four outlined types of contrastive patterning, in comparison with 43% judged to be unpatterned due to a variation involving one or more words. That is, the majority of all observed pen/voice use could be predicted by one of the identified contrastive functional patterns.

Simultaneous Mode Use

Simultaneously spoken and written input was rare in the present applications, with less than 0.5% of all words communicated in both modes. During the present studies involving verbal/temporal and computational/numeric content, however, 85% of content that was both spoken and written involved digits. In some cases, the digit primarily

References

- [1] P. Johnson, S. Feiner, J. Marks, M. Maybury, and J. Moore (eds.) *Intelligent Multimedia Multimodal Systems*. AAAI: Stanford University, March 1994. Working notes from Spring Symposium Series.
- [2] M. T. Maybury (ed.) *Intelligent Multimedia Interfaces*. AAAI Press/MIT Press: Menlo Park, California, 1993.
- [3] 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 J. Ohala (ed.), *Proceedings of the 1992 International Conference on Spoken Language Processing, vol. 2*, University of Alberta, October 1992, 1351–1354.
- [4] S. L. Oviatt, P. R. Cohen, M. Wang, and J. Gaston. A simulation-based research strategy for designing complex NL systems. In *ARPA Human Language Technology Workshop*, Morgan Kaufmann: San Mateo, California, March 1993.
- [5] S. L. Oviatt, P. R. Cohen, and M. Q. Wang. Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication*, December, 1994, in press.