Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction*

Sharon Oviatt, Antonella DeAngeli & Karen Kuhn**

Center for Human-Computer Communication

Department of Computer Science and Engineering

Oregon Graduate Institute of Science & Technology

P. O. Box 91000, Portland, OR 97291 USA

+1 503 690 1342

oviatt@cse.ogi.edu; http://www.cse.ogi.edu/~oviatt/

ABSTRACT

Our ability to develop robust multimodal systems will depend on knowledge of the natural integration patterns that typify people's combined use of different input modes. To provide a foundation for theory and design, the present research analyzed multimodal interaction while people spoke and wrote to a simulated dynamic map system. Task analysis revealed that multimodal interaction occurred most frequently during spatial location commands, and with intermediate frequency during selection commands. In addition, microanalysis of input signals identified sequential, simultaneous, point-and-speak, and compound integration patterns, as well as data on the temporal precedence of modes and on inter-modal lags. In synchronizing input streams, the temporal precedence of writing over speech was a major theme, with pen input conveying location information first in a sentence. Linguistic analysis also revealed that the spoken and written modes consistently supplied *complementary* semantic information, rather than redundant. One long-term goal of this research is the development of predictive models of natural modality integration to guide the design of emerging multimodal architectures.

Keywords

multimodal interaction, integration and synchronization, speech and pen input, dynamic interactive maps, spatial location information, predictive modeling

INTRODUCTION

As a new generation of multimodal/media systems begins to define itself, one theme that emerges frequently is the integration and synchronization requirements for combining different modes into a strategic whole system. From a linguistic perspective, the joint use of natural modes such as speech and manual gesturing has been described during human-human interaction, as has the role of gesture in both discourse and in thought [6,7]. Gesture has been viewed as a cognitive aid in the realization of thinking, and also as a carrier of different semantic content than speech:

"Speech and gestures are different material carriers... they are not redundant but are related, and so the necessary tension can exist between them to propel thought forward... to make the gesture is to bring the new thought into being on a concrete plane." [7, p.18]

The temporal synchrony between speech and gesture also has been analyzed for different languages [7,8].

Currently, little parallel work is available on modality integration during human-computer interaction, although such work will be crutial to guiding the design of planned multimodal systems. Using simulated systems, empirical research has begun to reveal that *contrastive functionality* is an influential theme in users' multimodal integration of speech and writing. That is, people use input modes in a contrastive manner to designate a shift in linguistic content or functionality-such as digit versus text, data versus command, or original versus corrected input [14,15]. Furthermore, during map-based tasks, interacting multimodally with speech and writing has numerous performance advantages over unimodal interaction, primarily because people have difficulty articulating spatial information [10]. In addition, users' frequency of composing multimodal commands is higher in visual/spatial domains than in verbal or quantitative ones [10]. Among other things, these data suggest that spatial domains may be ideal ones for developing early multimodal systems.

The purpose of this research was to conduct a comprehensive exploratory analysis of multimodal integration and synchronization patterns during pen/voice human-computer interaction. To achieve this, a simulation experiment was conducted in which people could combine spoken and pen-based input to interact multimodally while completing varied tasks using a dynamic map system. One goal of the study was to identify when users are most likely to compose their input multimodally, rather than unimodally. A task analysis of user commands was performed to distinguish the commonality of those expressed multimodally.

A second goal of this research was to analyze the main linguistic features of multimodal constructions, as well as differences from standard unimodal ones. Basic semantic constituents were examined to determine their content, order, and the preferred mode used to convey them. The type of pen input (e.g., graphics, symbols, pointing, words) also was analyzed for different types of multimodal task command.

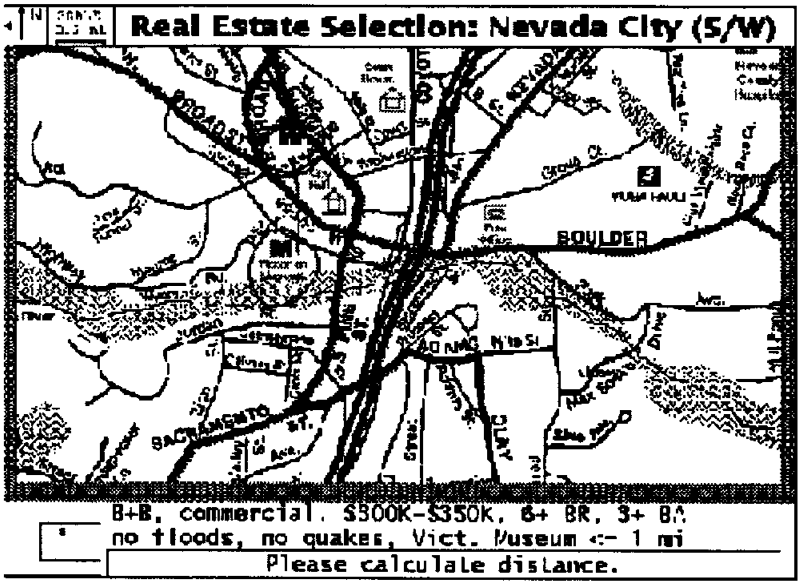
A third goal of this research was to investigate how spoken and written modes are naturally integrated and synchronized during multimodal constructions. The frequency of qualitatively different integration patterns was examined, such as sequential, simultaneous, and point & speak. Synchrony observed between the spoken and written signals was assessed for temporal precedence of one mode over the other, and for the typical lag between modes.

METHOD

Subjects, Tasks, and Procedure

Eighteen native English speakers participated in this research, half male and half female, and representing a broad spectrum of ages and professional careers.

A "Service Transaction System" was simulated that could assist users with map-based tasks. During a real estate selection task, participants were asked to select an appropriate home for a client. They were provided with a thumbnail sketch of the client's needs, such as acceptable price range. Using a city map, they filtered available homes until locating one meeting their constraints. For example, during such a task, a user could interact multimodally by circling a lakeside house icon with the pen and asking "Is this house in a flood zone? No flood zones, please." In response, the system would answer textually while displaying waterways and flood zones, and it would remove the house icon from the map if located in a hazard region. In a distance calculation, as shown in Figure 1, the user could circle two entities and connect a line between them while asking, "How far from here to here?" In response, the system would provide a numeric value in miles and a graphic confirmation of the map endpoints.



Real Estate Selection: Nevada City (S/W)

B+B, commercial, $300K-$350K, 6+ BR, 3+ BA
no floods, no quakes, Vict. Museum <:- 1 mi

Please calculate distance.

FIGURE 1. A multimodal distance calculation request, in which the user circles two locations and connects them with ink while speaking, "How far from here to here?"

During a map update task, people added, deleted, and modified information to represent changes in a high-growth municipal area. For example, a user could interact multimodally by drawing a square at a given location and saying "Make that a children's hospital." They also could draw a line along a road and say "Closed to traffic," or point to an arc across a highway and say, "Move this overpass here [drawing an arrow east] so the main hospital connects with the children's hospital." During all map tasks, users also controlled the map display by scrolling, automatically locating entities, zooming, and so forth, and they used speech and pen input for these controls too. In all cases, they interacted with an underlying map-based application as they added, removed, retrieved, or otherwise manipulated information to accomplish their task.

During the study, subjects received instructions, a general orientation to the map system's coverage, and practice using the system until its capabilities were clear. This orientation explained how to enter information on the LCD tablet when using the pen, speaking, or using a combination of both modes. During practice, users completed entire tasks using only speech or only pen, so they realized that the coverage of these alternative modes was equivalent. When writing, they were free to use cursive or printing, gestures, symbols, drawn graphics, pointing, or other marks. They were told to write information with the electronic stylus directly onto the color map displayed on their LCD tablet.

When speaking, subjects were instructed to tap and hold the stylus on the map as they spoke. A click-to-speak interface was used because off-line speech has been demonstrated to contain as many as 12,400% more unintelligible words than on-line speech directed to the system [13]. That is, massive differences can exist between the intelligibility and processability of speech in a click-to-speak versus open-microphone implementation, with click-to-speak interfaces presently offering the more viable alternative.

During the interactions reported in this study, people were free to use either or both input modes whenever they wished. They were encouraged to speak and write naturally, to work at their own pace, and to focus on completing their task. Since the goal was to uncover people's natural tendencies to interact multimodally and to integrate modes, an effort was made not to influence the manner in which they expressed themselves. People were told that the map system was well developed and tested, so it would be able to handle most of their input. If the system did not recognize their input, they always had the opportunity to re-enter their information.

People also were instructed on completing tasks using two different presentation formats: (1) a structured reference map, with the full network of roads, buildings, overlay information, and labels conventionally found on hard-copy reference maps, and (2) a less structured "minimalist" map, with one-third of the roads and overlay information as the more structured display, and only what was immediately needed to complete the task. Both map formats provided the same rapid interactivity and multimedia feedback (e.g., textual, graphic, synthetic speech) in response to user input.

After the session, a post-experimental interview was conducted in which users were asked their preferences and evaluation of the system. All users reported believing that the "system" was a functional one, after which they were debriefed about simulation details.

Semi-Automatic Simulation Technique

People's input actually was received by an informed assistant, who performed the role of interpreting and responding as the system. The assistant tracked the subject's written and spoken input, and clicked on predefined fields at a Sun SPARCstation to send new map displays and confirmations back to the subject. An emphasis was placed on automating the simulation to create rapid subject-paced interactions (i.e., averaging less than 1 sec delay) with clear feedback. Details of the simulation method, capabilities, environment, and performance characteristics have been provided elsewhere [11], although the method was adapted extensively for this study to handle the dynamic display of maps, overlays, and photographs.

Research Design and Data Capture

Each of the 18 subjects completed one real estate and one map update task in each of the two formats-or four map tasks apiece. Therefore, the present analyses focus on data collected during 72 tasks, half using each of the two map formats, with order counterbalanced.

All interaction was videotaped and included a real-time record of all spoken and written input and system responses. Hardcopy multimodal transcripts also were created, with the subject's written input captured automatically in the current map context, and verbatim spoken input transcribed onto the printouts. Sequencing information was annotated for the two input streams, including temporal overlap at the word level.

Transcript Preparation and Coding

Coding was conducted on the multimodal corpus for the following dependent measures:

User Preference

The percentage of subjects who chose to interact unimodally or multimodally during map tasks was summarized, as was the percentage of subjects who reported during interviews that they preferred to interact either unimodally or multimodally.

Task Actions

Individual user commands to the map system were classified into the following types of action command: (1) *Add* object or subset of objects-which could involve specifying spatial location information about a point(s), line, or area-e.g., "Add historic homes here"; "Add open space park," (2) *Move* object to new location-e.g., "Move highway on-ramp to here," (3) *Modify* specific route or spatial area- e.g., "Close road west of May Lake as shown," (4) *Calculate distance* between two locations-e.g., "How far from here to here?" (5) *Query for information* about object-e.g.,"Is this house in Nevada City School district?" (6) *Delete* object-e.g., "Erase this line," (7) *Label* object-e.g., "This is an apple orchard," (8) *Zoom* on object-e.g., "View house," (9) *Control task* procedures-e.g., "Next task please," (10) *Scroll* map-e.g., "Let's go up this way," (11) *Print* screen display-e.g., "Print photo," (12) *Automatically locate* out-of-view object-e.g., "Show me American Hill Park," (13) *Call up overlay*-e.g., "Show lakes," and (14) *Specify constraints* for filtering information-e.g., "Show me houses between $300,000 and $350,000." Each individual user command found in the corpus was classified by type, and was scored as having been expressed unimodally or multimodally.

Linguistic Content

Analyses were conducted of multimodal constructions to determine whether speech or writing was used to convey each semantic constituent. To assess whether the input order of multimodal constructions conformed with the canonical S-V-O ordering expected for English, the order of basic semantic constituents such as subject (S), verb (V), object (O), and locatives (LOC, e.g., "east of May Lake") was analyzed for individual multimodal constructions and unimodal spoken ones.

For pen-based input, the basic type of semantic content was classified into the following: (1) drawn graphics (e.g., rectangle to indicate a building), (2) symbols and signs (e.g., > to indicate greater than), (3) simple pointing, (4) full and abbreviated lexical words (e.g., BR for bedroom), and (5) digits. The percentage of total pen input representing each of these categories was summarized for multimodal constructions as a function of task command, and for unimodal written ones.

The type and frequency of spoken spatial deictics (e.g., "there") was summarized, as was the percentage of multimodal constructions containing a deictic term.

Multimodal Integration Patterns

Individual constructions were classified into these integration patterns: (1) Simultaneous-spoken and written input overlapped temporally, (2) Sequential-either spoken or written input preceded, with the other mode following after a time lag, (3) Point & speak-pointing at or on the border of an object while simultaneously speaking about it, but without creating drawn marks other than a singular dot, and (4) Compound-a two-part sequence involving a written input phase and a point & speak phase (e.g., creation of drawing, then point-and-speak about it).

For the sequential and compound integration patterns, microanalyses were conducted from videotapes of which mode preceded the other, and of the average time lag between the end of the first input mode and the onset of the second one. Simultaneous constructions were classified into nine logically possible overlap patterns, displayed in Table 4. These classifications were designed to code the relative temporal order of signal onset and offset for spoken and written input, and to provide temporal distinctions about coordination between signals accurate to within 0.1 sec. Simple point & speak input was not included in this analysis of simultaneity, since it was considered to involve completely overlapped signals by necessity of the click-to-speak interface.

In addition to analyses at the utterance level, the integration of spoken and written input was analyzed for multimodal constructions with a spoken deictic. In these cases, the temporal relation between the spoken deictic term and the specific pen-based mark that disambiguated the deictic's meaning were microanalyzed to determine whether they occurred simultaneously or sequentially, and to assess typical precedence relations and time lags. These analyses were based on temporal information about the onset and offset of the spoken deictic term, as well as spatial/temporal information about the beginning and end of the formation of the relevant written mark, which were analyzed from videotapes.

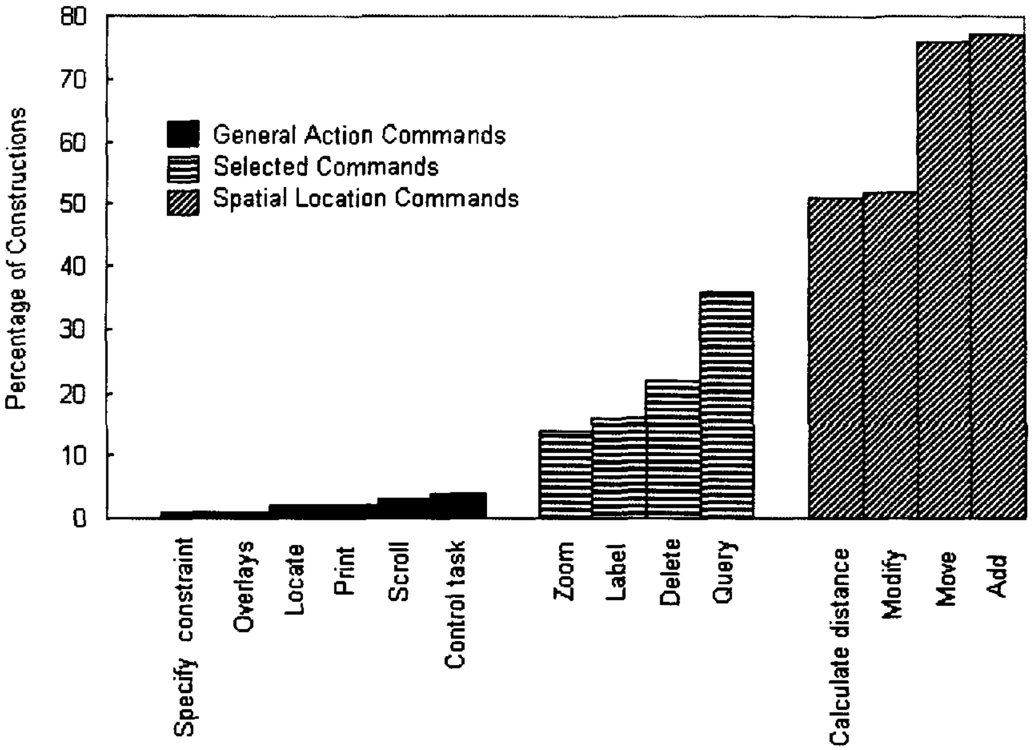


FIGURE 2. Percentage of all constructions that users expressed multimodally as a function of task command, with spatial location commands on the right, selection commands in the middle, and general action commands on the left.

Reliability

All reported measures had reliabilities of 0.80 or above, with inter-modal lags accurate to within 0.1 sec.

Implemented System Testing

Additional testing was conducted to determine whether multimodal integration patterns would remain the same with a: (1) click-to-speak vs. open-microphone interface implementation, and (2) simulated vs. actual system testing. The Quickset system, which recognizes spoken and pen-based input to maps and has been modeled on the present simulation [16], was adapted to accept open-microphone speech for testing purposes. Data then were collected from 8 subjects as they provided spatial location commands (e.g., add or move objects) to the Quickset system.

RESULTS

Since no differences were found in any measures due to the map's visual display, this section reports results collapsed over format.

User Preference

Users had a strong preference to interact multimodally during map tasks, rather than unimodally. All of them, or 100%, used both spoken and pen input at different points during each task. During interviews, 95% of users also reported a preference to interact multimodally.

Of 871 individual constructions, users expressed 167 multimodally by combining speech and writing within the same sentence, or 19%. Unimodal writing accounted for 17.5% of all sentences, and the remaining 63.5% of sentences were uttered just using speech.

Task Action Analysis

Figure 2 illustrates that type of task command greatly influenced the likelihood of expressing an utterance multimodally. In particular, the four user commands most likely to be expressed multimodally were ones that required specifying a spatial location description. These *spatial location commands,* which accounted for 86% of sentences that users constructed multimodally, are displayed on the right side of Figure 2. They tended to involve graphic input to add, move, modify, or calculate the distance between objects. Spatial location commands were the only ones more likely to be expressed multimodally than unimodally-ranging between 51 and 77% of the time.

Commands that involved selecting a specific object from others displayed on the map had an intermediate likelihood of being expressed multimodally. These *selection commands* accounted for 11% of users' multimodal constructions, ranging between 14 and 36% (Figure 2, middle). Such commands identified an object of interest and its location, but no complex spatial information. They included querying for information about an object, and deleting, labeling, or zooming on an object. Commands that involved selecting an in-view object were more likely to be expressed unimodally than multimodally, because the object sometimes was already in focus from: (1) previous dialogue context (e.g., user adds new map object, then deletes it), or (2) visual context (e.g., user zooms on a house photo, then queries for information about it). Sometimes the object simply was one of a kind or easy to describe. In none of these cases did the user have a compelling need to physically gesture to an object to select it.

The remaining six types of task command were rarely expressed multimodally, accounting for only 3% of all constructions (Figure 2, far left). This third subgroup involved *general action commands,* which required neither a spatial description nor identification of an in-view object. They included controlling task procedures, scrolling the display, printing, automatically locating out-of-view objects, calling up map overlays, and specifying constraints for filtering information.

A Wilcoxon Signed Ranks analysis confirmed that spatial location commands were significantly more likely to be expressed multimodally than selection commands, T+ = 115 (df = 15), $p < .0003$, one-tailed. In addition, selection commands were significantly more often multimodal than general action commands, T+ = 28 (df = 7), $p < .008$, one-tailed.

Linguistic Content

Multimodal constructions were expressed in a telegraphic command language, which was briefer and less complex syntactically than unimodal spoken sentences (see [12] for full linguistic analysis and processing implications). The following illustrate typical user input from transcripts.

Unimodal speech:

"Add a boat dock on the west end of Reward Lake."

"I want to see the photo of the house on the southwest end of Reward Lake, please."

Multimodal:

[draws line] "Add dock here."

[circles house] "Show photo."

Order of Semantic Constituents

As expected, 98% of the unimodal spoken constructions conformed with the standard subject-verb-object order typical of English. Almost all, or 97%, of multimodal constructions also conformed with this expected order, and when basic constituents were elided the remaining ones still conformed (e.g., V-O, as in "Show photo"). The primary difference between spoken and multimodal sentences, as Table 1 clarifies, was in the typical position of locative constituents. Spoken utterances rarely began with a locative descriptor or clause (i.e., 1% of constructions), and instead reserved LOCs for sentence-final position (i.e., 96%)-as in "Add apple orchard east of Sugarloaf Mountain Park" [V-O-LOC]. In contrast, multimodal constructions invariably began with drawn graphics conveying LOC information (i.e., 95%) and were followed by spoken S-V-O constituents-as in [draws oval] "Add pool" [LOC-V-O], and LOCs never occurred in sentence-final position.

TABLE 1. Percentage of multimodal and speech-only constructions for which the locative constituent [LOC] occurred in sentence initial vs. final position, rather than mid-sentence.

	Initial LOC	Final LOC
Speech-only	1%	96%
Multi-Modal	95%	0%

Mode of Semantic Constituents

In multimodal constructions, pen input was used 100% of the time to convey location and spatial information about objects (i.e., size, shape, number). In 2% of cases, speech provided duplicate but less precise information about location constituents. In comparison, speech was used for 100% of subject and verb constituents, although most subjects were elided as a result of the command language style. The majority of object constituents, or 85%, also were spoken, although in 15% of cases written input identified the specific object. At a semantic level, then, speech and writing clearly contributed different and *complementary information.* It was rare for information to be duplicated in both modes. However, there is a sense in which spoken deictics provided duplicate information, since they simply flagged the fact that a location or object had been indicated in writing. Likewise, drawn graphics provided partially duplicated information about the type of object being added to the map (e.g., rectangles to indicate a shopping center), although subjects always followed drawings with more precise spoken object descriptions.

Type of Pen-based Content

Of the pen-based written input that people used during multimodal constructions, the majority, or 48%, involved drawn graphics (e.g., a square to represent a building; a line to represent roadway). Another 28% involved symbols or signs (e.g., an X to delete; an arrow to indicate movement), and 17% involved simple pointing. Only 7% of written input were actual words, and none were digits.

Table 2 illustrates that different classes of written input predominated during different types of multimodal task command. Graphic input was the most prevalent during spatial location commands, occurring primarily when an object was added to the map. In contrast, pointing predominated during selection commands, and written words during general action commands. Written symbols and signs were used in a relatively stable manner, irrespective of command type. For comparison, the far right column of Table 2 illustrates that words were predominant during unimodal written sentences, with 52% of these general action commands.

TABLE 2. Categories of pen-based input during multimodal constructions (listed by task command), and unimodal written constructions.

	Multimodal Constructions by Command Type			Unimodal Pen Constructions
	General Action	Selection	Spatial Location	
Graphic	0%	9.5%	53%	9%
Symbol	25%	33.5%	27.5%	32%
Pointing	0%	57%	14%	0%
Words	75%	0%	5.5%	48%
Digits	0%	0%	0%	11%

Deictic Terms

Most multimodal utterances, or 59%, did *not* contain a spoken deictic term. When present, 96% of deictics involved the terms "here," "there," "this," or "that" (e.g., user circles house and says: "Is this brick?").

Multimodal Integration Patterns

Table 3 reveals that 86% of the 167 multimodal constructions involved a draw and speak pattern, with *simultaneous* integration of drawing and speaking in 42% of constructions, *sequential* input in 32%, and a *compound* pattern in another 12% (i.e., draw graphic, then point to it while speaking). A *point and speak* pattern occurred far less frequently than drawing and speaking- in just 14% of constructions.

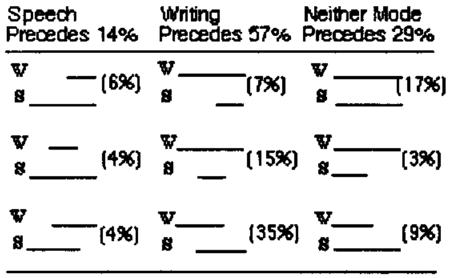
TABLE 3. Percentage of multimodal constructions represented by different types of speech/writing integration pattern.

Type of Integration Pattern	Percent of Multimodal Constructions
Point & Speak	14%
Draw & Speak	86%
---Simultaneous Draw & Speak	(42%)
---Sequential Draw & Speak	(32%)
---Compound Draw & Speak	(12%)

Simultaneous Integrations

Simultaneous draw-and-speak constructions were classified into nine possible synchronization patterns, displayed in Table 4 with the percent of simultaneous constructions represented by each. Constructions in which the speech signal showed temporal precedence (left column) accounted for 14% of the total, whereas those in which writing preceded speech (middle column) accounted for most of the total, or 57%. In the remaining cases, neither mode preceded (right column). A Wilcoxon Signed Ranks analysis confirmed that written input was significantly more likely to precede speech than the reverse, T+ = 42 (N = 9), p < .01, one-tailed.

TABLE 4. All logically-possible temporal overlap patterns between speech and written input for simultaneous integrations, subclassified by temporal precedence of input mode.



Two subjects produced utterances in which speech was abnormally elongated as a result of attempting to perfectly synchronize the beginning and end of speech and drawing. For example, while marking a closed section of road one subject said, "No automobiles" (underlined syllables elongated). However, only 2% of multimodal utterances were affected by such distortion.

Sequential Integrations

Analysis of the sequential constructions again revealed the temporal precedence of written input. In 99% of such constructions, a drawn graphic was completed before the onset of spoken input. A Wilcoxon Signed Ranks analysis confirmed that users were significantly more likely to complete their pen input before speaking than the reverse, T+ = 105 (N = 14), p < .001, one-tailed.

Figure 3 illustrates the distribution of lags for sequential constructions. The lag between the end of the pen signal and start of speech averaged 1.4 secs, with 70% of all lags ranging between 0.0 and 2.0 sec, 88% between 0.0 and 3.0 sec, and 100% between 0.0 and 4.0 sec.

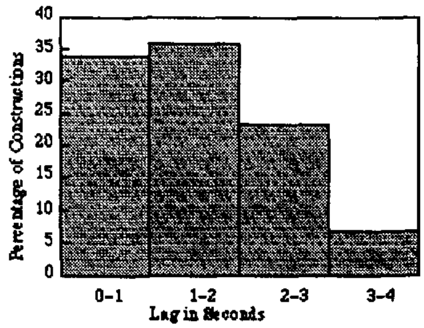


FIGURE 3. Distribution of lag times between end of pen signal and onset of speech in sequential multimodal constructions.

Integration of Deictic Terms

In addition to analyses at the utterance level, the temporal relation between spoken deictic terms and the specific written mark that disambiguated the deictic's meaning also was analyzed to determine whether these related signals co-occurred. Out of the 41% of multimodal constructions that *did* contain a spoken deictic, most (43%) displayed a *sequential* integration pattern. In 100% of these cases, the drawn mark that disambiguated the deictic's meaning was completed *before* the deictic term was spoken, with an average lag of 1.1 sec and a maximum lag of 3 sec 97% of the time. In another 27% of multimodal constructions, the spoken deictic and relevant ink occurred *simultaneously*. Further analysis of the temporal overlap revealed that the onset of writing still preceded speech in 60% of these cases, and in only 5% did speech onset precede writing. In the remaining 30% of multimodal constructions, spoken deictics were accompanied by simultaneous *pointing* or a *compound* integration pattern. In summary, the temporal precedence of written input continued as a theme in deictic integration patterns.

Integration in Open-Microphone System

Testing with an open-mic system was conducted mainly to evaluate whether use of the pen to engage the click-to-speak mechanism might influence the temporal precedence of written input over speech. However, the results of analyses on 170 additional multimodal constructions replicated simulation findings: (1) in sequential constructions, pen input preceded speech 100% of the time, and (2) in simultaneous constructions, the onset of pen input preceded speech 61% of the time (speech preceded 5%; neither mode 34%). The general frequency of sequential and simultaneous constructions also was comparable to that found in the simulation data, and pen input again conveyed locative semantic content in LOC-S-V-O order.

DISCUSSION

Although users overwhelmingly preferred to interact multimodally rather than unimodally, they nonetheless did not issue every command to the system multimodally. Knowledge of the type of task command provides considerable predictive information about its likelihood of being expressed multimodally. During interaction with complex visual displays such as maps, *spatial location commands* (e.g., adding, moving) were by far more likely to be composed multimodally than unimodally-by a factor of 2-to-1. In fact, spatial location commands accounted for the vast majority, or over 86%, of the sentences that users chose to express multimodally. For these commands, users needed to specify information about the location, number, size, and/or shape of a point, line, or area. In contrast, *selection commands* in which users identified one object from a set (e.g., zooming on, deleting) only occasionally were expressed multimodally, since the presence of a unique descriptor, immediate dialogue context, or the map context often made physical selection unnecessary. In users' view, when contextual information was present, they expected to use it and did not automatically issue a selection gesture too. Likewise, *general action commands* that entailed neither spatial nor selection information (e.g., calling up overlay, printing) rarely were expressed multimodally-only 1% of the time.

One implication of these findings is that knowledge of the task commands anticipated in an application could influence the fundamental design choice to build a multimodal versus unimodal interface. In a multimodal system, knowledge of a given command (generally indicated by the spoken verb) also could be used to weight likelihoods that the incoming signal is: (1) unimodal, or (2) part of a multimodal construction in which speech input is expected to follow pen within a given lag. In the latter case, knowledge about the type of command could influence architectural decisions about when to begin processing, and the signal's interpretation.

Among the powerful interface features of pen-based input are its ability to convey precise location information and detailed graphic renderings. Another is the multi-functional capability of pen systems, which can support qualitatively different types of input such as drawings, symbols and signs, gestures, words, digits, and pointing. The present data indicate that knowledge of the command type also provides predictive information about the kind of pen input most likely to be elicited from users, which will need to be processed by the recognizer. For example, spatial location commands (e.g., add) primarily elicited drawn graphics, whereas selection commands elicited pointing and gestures (e.g., circling an object), and general action commands elicited words. In designing future multimodal applications, information about expected task commands therefore ought to be considered before specifying a planned system's basic recognition capabilities.

The past literature on multimodal systems has focused largely on simple selection of objects or locations in a display, rather than considering the broader range of multimodal integration patterns. In this corpus, speech combined with pointing for selection was *not* the dominant integration theme, accounting for only 17% of multimodal constructions. Most pen input was not written words either (accounting for 7%), perhaps contrary to expectations of the handwriting recognition community. Instead, drawn graphics (e.g., square for building) and symbols/signs (e.g., arrow to indicate movement) accounted for most, or 76%, of all written input. Given the more powerful and multifunctional capabilities of new pen devices, which can generate symbolic information as well as selecting things, it is clear that a broader set of multimodal integration issues needs to be addressed in future work.

Previous specialized processing approaches based on the interpretation of spoken deictics via synchronous pointing (i.e., by "calling out" for a matching x,y coordinate on a display to resolve an intended referent in a phrase like "that blue square" [1,4]) are unlikely to play a large role in handling the types of construction actually observed in the present corpus. First, most multimodal constructions, or 59%, did *not* contain any spoken deictic, so one cannot count on their presence to flag and assist in interpreting the referent in a visual display. Second, even fewer multimodal constructions, or 25%, contained a spoken deictic that overlapped in time with the pen input needed to disambiguate its meaning. Third, as noted above, only 17% of multimodal constructions involved a simple point-and-speak pattern. Finally, as the present data attest, users actually may only compose individual sentences multimodally a limited percentage of the time. To process what may be as many as 80% of linguistic constructions unimodally in a multimodal-capable interface, a system designed for a real application must be able to interpret standard unimodal referring expressions and resolve reference through both dialogue and visual context as in previous multimodal designs [2,9]. In this context, specialized algorithms for processing deictic-point relations have limited practical utility.

One important distinguishing characteristic of spoken and pen input is that both modes can convey symbolic content such as language. Analysis of the linguistic content of integrated speech/writing constructions in this study revealed several interesting things. First, at a semantic level, the spoken and written modes consistently contributed different and *complementary information*. Basic constituents describing the subject, verb, and object almost always were spoken, whereas constituents describing locative information invariably were written. Furthermore, consistent with McNeill's [7] observations, it was extremely rare for such information to be duplicated in both modes. These data confirm the importance of *contrastive functionality* as a major theme that drives the overall patterning of people's integrated use of input modes [14]-with locative/nonlocative content the salient contrast in this visual/spatial domain. Second, multimodal constructions were briefer and syntactically simpler than unimodal spoken ones, and therefore potentially easier for a system to process (see [12] for further discussion). Third, the order of incoming linguistic information in multimodal constructions clearly departed from the canonical S-V-O-LOC order typical of spoken English. Instead, pen-based locative information was presented first and followed by spoken constituents, resulting in a LOC-S-V-O sequence.

With respect to synchronization of input streams, a major theme for both sequential and simultaneous patterns was the strong temporal precedence of written input, which prevailed independent of the click-to-speak or open-microphone implementation. During sequentially integrated draw-and-speak constructions, a drawn graphic was completed before the onset of any spoken input 99% of the time. Analysis of the lags revealed that speech followed writing within an average of 1.4 seconds, and always began within 4 seconds of pen input. When drawing and speech overlapped in simultaneous constructions, the onset of pen input still preceded speech more often than the reverse (57% vs. 14% of cases). Finally, analysis of spoken deictics and their disambiguating marks revealed that pen input preceded the deictic term 100% of the time when these signals were sequential, and 60% of the time when simultaneous. This observed precedence of pen input generalized over both system and simulation testing, involving click-to-speak and open-microphone interfaces. Future simulation research should explore typical integration patterns between other promising modality combinations, such as speech and 3-D gestures or speech and gaze, for interacting with other types of visual display- as well as their relation to the spoken and pen-based integration patterns reported here.

One interpretation of the temporal precedence of writing to convey locative content is that users were elaborating the visual context of the map with their ink marks and, after this expanded context was available, they then continued by speaking about it. The act of drawing and permanence of the written marks may have had an important self-organizing influence on users thinking and subsequent speech. During interpersonal communication, both signed language and natural gestures also have been reported to precede or occur simultaneously with their spoken lexical analogues [3,5,8]. Some variation has been found in integration patterns between languages, such that topic-prominent languages like Chinese present gestures further in advance of the speech stream (i.e., as a kind of "framing constraint" for the sentence) than do subject-prominent languages like Spanish or English [7]. Although gesturing is ephemeral and seemingly unlike the permanence of ink, people sometimes engage in a "poststroke hold" that can perpetuate the gesture as a visual context for speech in the same way that ink does. In this sense, the dynamics of context-setting can function similarly in eliciting advance writing and manual gesturing.

From a more pragmatic perspective, the order of input modes and average lag times reported in this paper could be used to weight probabilities associated with the likelihood that a sentence is multimodal versus unimodal, the likelihoods associated with different utterance segmentations (e.g., that an input stream containing [speech, writing, speech] should be segmented into [S / W S] rather than [S W / S]), and to correctly recognize content within the spoken and written input streams. Current systems that time-stamp and jointly process two or more input modes have not reported temporal thresholds for performing integrations between modes. Data on typical inter-modal lags collected during realistic interactive tasks, such as those reported here, could form the basis of highly accurate mode integrations in future multimodal systems.

The present empirical research has inspired the design and architectural implementation of multimodal systems in our laboratory, which support map-based applications ranging from real-estate and health-care selection to military simulation [16]. In these systems, the user communicates through a hand-held PC that processes speech and pen input in parallel, using a joint interpretation strategy involving a statistically-ranked unification of semantic interpretations. Compared with unimodal recognition, such systems have the advantage of supporting mutual disambiguation of linguistic content and reduction of error. Given the complex and nonintuitive nature of users' multimodal interaction during real tasks, empirical work will be essential in guiding the design of future robust multimodal systems.

FOOTNOTES

* This research was supported in part by NSF Grants No. IRI-9213472 and IRI-9530666, DARPA Contract DABT63-95-C-0007, and grants, contracts, and equipment donations from Apple, AT&T/NCR, ETRI, Intel, NTT, Scriptel, Southwestern Bell, Sun Microsystems, and US West.

** Collaborators' affiliations: Psychology Dept., University of Trieste, and Linguistics Dept., Portland State University.

[1] Empirical analysis confirmed that intentional pointing to a particular referent was distinct from untargeted tapping on the tablet simply to engage the click-to-speak interface (i.e., for which the pen could drop to the nearest tablet location), with the former averaging 1.7 sec, versus 1.4 sec for the latter.

[2] Spoken deictic terms such as "here" and "this" point out locations in the spatial context shared by communication participants, and often are accompanied by gesturing.

[3] Gesture formation can be classified into a preparatory phase, main stroke, poststroke hold, and retraction.

ACKNOWLEDGMENTS

Thanks to A. Cheyer, M. Reinfreid, and L. Waugh for assistance with programming the simulation and acting as simulation assistant, and to P. Schmidt and D. McGee for preparing the open-mic Quickset system. Thanks also to P. Li, E. Olsen, M. Heitz and R. vanGent for transcription and data analysis, and to P. Cohen, M. Johnston and a CHI reviewer for valuable comments.

REFERENCES

1. Bolt, R. "Put-That-There": Voice and Gesture at the graphics interface, *Computer Graphics*, 1980, 14 (3): 262-270.

2. Cohen, P., Dalrymple, M., Moran, D. & Pereira, F. Synergistic use of direct manipulation and natural language, *CHI '89 Conf. Proc.,* ACM: Addison Wesley, New York, 1989, 227-234.

3. Kendon, A. Gesticulation and speech: Two aspects of the process of utterance, *The Relationship of Verbal and Nonverbal Communication* (ed. by M. Key), The Hague: Mouton, 1980, 207-227.

4. Koons, D., Sparrell, C. & Thorisson, K. Integrating simultaneous input from speech, gaze, and hand gestures, *Intelligent Multimedia Interfaces*, ed. by M. Maybury, MIT Press: Cambridge, MA, 1993, 257-76.

5. Levelt, W., Richardson, G. & Heu, W. Pointing and voicing in deictic expressions, *Jour. of Memory and Language,* 1985, 24, 133-164.

6. McNeill, D. Hand and Mind: What gestures reveal about thought, Univ. of Chicago Press: Chicago, Ill., 1992.

7. McNeill, D. Language as gesture (Gesture as language), *Proc. of the Workshop on the Integration of Gesture in Language & Speech*, ed. by L. Messing, Univ. of Delaware, Oct. 1996, 1-20.

8. Naughton, K. Spontaneous gesture and sign: A study of ASL signs co-occurring with speech, *Proc. of the Workshop on the Integration of Gesture in Language & Speech*, ed. by L. Messing, Univ. of Delaware, Oct. 1996, 125-34.

9. Neal, J. & Shapiro, S. Intelligent multi-media interface technology, in *Intelligent User Interfaces* (J. Sullivan & S. Tyler, eds.), ACM: Addison Wesley, New York, 1991, ch. 3, 45-68.

10. Oviatt, S.L. Multimodal interfaces for dynamic interactive maps, *CHI '96 Conf. Proc.,* New York, ACM Press, 1996, 95-102.

11. Oviatt, S., Cohen, P., Fong, M., & Frank, M. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting, *Proc. of the Intl. Conf. on Spoken Language Processing,* 1992, 2, 1351-54.

12. Oviatt, S., Cohen, P. Johnston, M. & Kuhn, K. Multimodal language: Linguistic features and processing requirements, forthcoming.

13. Oviatt, S., Cohen, P. & Wang, M. Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity, *Speech Communication*, 1994,15 (3-4), 283-300.

14. Oviatt, S. & Olsen, E. Integration themes in multimodal human-computer interaction, *Proc. of the Intl Conf. on Spoken Language Processing,* 1994, 2, 551-554.

15. Oviatt, S. L. & vanGent, R. Error resolution during multimodal human-computer interaction, *Proc. of the Intl. Conf. on Spoken Language Processing*, 1996.

16. Pittman, J., Cohen, P., Smith, I., Yang, T. & Oviatt, S. Quickset: A multimodal interface for distributed interactive simulations, *Proc. of the 6th Conf. on Computer-Generated Forces & Behavior Representation*, Univ. of Central Florida, Orlando, FL., 1996, 217-24.