Toward a Theory of Organized Multimodal Integration Patterns during Human-Computer Interaction

Sharon Oviatt[1] Rachel Coulston[1] Stefanie Tomko[2] Benfang Xiao[1]
Rebecca Lunsford[1] Matt Wesson[1] Lesley Carmichael[3]

[1]Oregon Health & Science Univ.
OGI School of Sci. & Eng.
20000 NW Walker Road
Beaverton, OR 97006, USA
+1 503 748 1342
oviatt@cse.ogi.edu
{rachel, benfangx,
rebeccal}@cse.ogi.edu

[2] Carnegie Mellon University
Language Technologies Institute
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
+1 412 268 9503
stef@cs.cmu.edu

[3]University of Washington
Department of Linguistics
Box 354340
Seattle, WA 98195-4340, USA
+1 206 543 2046
lesley@u.washington.edu

ABSTRACT

As a new generation of multimodal systems begins to emerge, one dominant theme will be the integration and synchronization requirements for combining modalities into robust whole systems. In the present research, quantitative modeling is presented on the organization of users' speech and pen multimodal integration patterns. In particular, the potential malleability of users' multimodal integration patterns is explored, as well as variation in these patterns during system error handling and tasks varying in difficulty. Using a new dual-wizard simulation method, data was collected from twelve adults as they interacted with a map-based task using multimodal speech and pen input. Analyses based on over 1600 multimodal constructions revealed that users' dominant multimodal integration pattern was resistant to change, even when strong selective reinforcement was delivered to encourage switching from a sequential to simultaneous integration pattern, or vice versa. Instead, both sequential and simultaneous integrators showed evidence of *entrenching* further in their dominant integration patterns (*i.e.*, increasing either their inter-modal lag or signal overlap) over the course of an interactive session, during system error handling, and when completing increasingly difficult tasks. In fact, during error handling these changes in the co-timing of multimodal signals became the main feature of hyper-clear multimodal language, with elongation of individual signals either attenuated or absent. Whereas Behavioral/Structuralist theory cannot account for these data, it is argued that Gestalt theory provides a valuable framework and insights into multimodal interaction. Implications of these findings are discussed for the development of a coherent theory of multimodal integration during human-computer interaction, and for the design of a new class of adaptive multimodal interfaces.

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (HCI)**]: User Interfaces – *user-centered design, theory and methods, interaction styles, input devices and strategies, evaluation/methodology, voice I/O, natural language, prototyping.*

General Terms

Performance, Design, Reliability, Human factors

Keywords

Multimodal integration, speech and pen input, co-timing, entrenchment, error handling, task difficulty, Gestalt theory

1. INTRODUCTION

As a new generation of multimodal systems begins to emerge, one dominant theme will be the integration and synchronization requirements for combining modalities into robust whole systems. In this respect, the design of future multimodal systems depends critically on accurate knowledge of the natural integration patterns that typify people's combined use of different input modes. Currently, the two most mature types of multimodal interface are ones that jointly process speech and pen input or speech and lip movement input [1, 10]. Within each of these areas, research has examined various aspects of the temporal synchronization patterns that characterize multimodal integration [5, 6, 12]. This paper will describe quantitative modeling on the organization of users' speech and pen multimodal integration patterns, as well as present a theoretical context for interpreting these new data.

1.1 Multimodal Integration Myths

One "myth" that misguided computationalists' thinking about early multimodal system design has been the belief that users' multimodal input always involves simultaneous signals [9]. More recently, empirical evidence has clarified that multimodal constructions frequently do not co-occur temporally during either human-computer or natural human communication. Rather, multimodal input often is integrated sequentially, with a

manually-oriented input mode (*e.g.*, natural or signed manual gestures, pen input) typically delivered earlier than speech input [7, 8, 12]. In the case of speech and pen, pen input frequently precedes speech with a brief lag between input modes of 1-2 seconds [12]. Ideally, future multimodal systems should not only be able to process both unimodal and multimodal user input, but also both *simultaneously* and *sequentially* integrated multimodal constructions.

Another "myth" has been the belief that all users' multimodal input is integrated in a uniform way [9]. To the contrary, recent empirical evidence on multimodal speech and pen interaction has revealed an unusual bimodal distribution of user integration patterns. As illustrated in Figure 1, previous data indicate that individual child, adult, and elderly users all adopt either a predominantly simultaneous or sequential integration pattern during speech and pen multimodal constructions [9, 14, 15]. In these studies, users' dominant integration pattern was identifiable almost immediately, typically on the very first multimodal command, and remained highly consistent (88-93%) throughout a session. These findings imply that future multimodal systems that can detect and *adapt to a user's dominant integration pattern* potentially could yield substantial improvements in system robustness.

Children			Adults			Seniors		
User	**SIM**	**SEQ**	**User**	**SIM**	**SEQ**	**User**	**SIM**	**SEQ**
SIM integrators:			**SIM integrators:**			**SIM integrators:**		
1	100	0	1	100	0	1	100	0
2	100	0	2	94	6	2	100	0
3	100	0	3	92	8	3	100	0
4	100	0	4	86	14	4	97	3
5	100	0	**SEQ integrators:**			5	96	4
6	100	0	5	31	69	6	95	5
7	98	2	6	25	75	7	95	5
8	96	4	7	17	83	8	92	8
9	82	18	8	11	89	9	91	9
10	65	35	9	0	100	10	90	10
SEQ integrators:			10	0	100	11	89	11
11	15	85	11	0	100	12	73	27
12	9	91				**SEQ integrators:**		
13	2	98				13	1	99
						Non-dominant integrators:		
						14	59	41
						15	48	52
Average Consistency 93.5%			**Average Consistency 90%**			**Average Consistency 88.5%**		

Figure 1. Percentage of simultaneously-integrated multimodal constructions (SIM) versus sequentially-integrated constructions (SEQ) for children, adults, and seniors

Engineering-level concepts tend to dominate whenever new multimodal systems are built, even though it now is widely recognized that well designed multimodal systems depend critically on guidance from cognitive science, linguistics, and other areas. As a new class of *adaptive multimodal interfaces* begins to be prototyped, engineers might reasonably ask whether users can't just be trained to deliver their multimodal commands in a simultaneously integrated manner. This could expedite the multimodal fusion process and, in particular, could simplify the development of temporal constraints that are needed to build new time-sensitive multimodal architectures. The present research explores this theme of the potential malleability of users' multimodal integration patterns, as well as examining variation in users' integration patterns during system error handling and tasks varying in difficulty.

1.2 Behavioral/Structural vs. Gestalt Theory

As applied to multimodal communication and integration patterns, Behavioral/Structuralist theory and Gestalt theory each would predict a strikingly different if not an opposite pattern of results. A Behavioral/Structuralist perspective would view multimodal communication as being composed of a set of discrete signal pieces (*e.g.*, speech and pen), and also governed by the same principles that apply to these individual modalities. Likewise, a Behavioral/Structuralist approach would assert that an individual's multimodal integration pattern should be malleable, or subject to external conditioning via standard stimulus-response training techniques.

In contrast, Gestalt theory would describe multimodal integration patterns as forming a unique or qualitatively different whole form, one that transcends the simple sum of their individual signal parts. The Gestalt viewpoint originally was developed as a revolt against both the Structuralist and Behaviorist theories that analyzed experience into simpler components. Literally hundreds of illustrations have been documented in which Gestalt patterns are perceived that otherwise would have been overlooked if complex patterns had been artificially fragmented or decomposed into single stimuli. One classic example is the case of Wertheimer's demonstration in 1912 that two lines flashed successively at optimal intervals appear to move together, an illusion related to human perception of motion pictures [3]. Although Gestalt theory's main contributions have involved elucidation of human perception of visual-spatial phenomena, they also have been applied to the perception of acoustic, haptic, and other information [2].

One of Gestalt theory's contributions has been the description of principles for grouping information into a coherent whole [3, 4]. For example, the *principle of proximity* states that spatial or temporal proximity causes elements to be perceived as related. In a multimodal pen/voice interaction, speech is an acoustic modality that is structured temporally, and pen input is a visible modality that is structured both spatially and temporally. In this case, Gestalt theory would predict that the *common temporal dimension* would provide organizational cues for binding these modes into a whole multimodal communication. That is, changes in the co-timing of these modalities could well serve as the main information conveying proximity and relatedness.

In addition, the Gestalt *principle of symmetry* states that people have a tendency to perceive symmetrical elements as part of the same whole. In the case of pen/voice multimodal communication, a symmetrical arrangement would entail closer temporal correspondence between signal pieces, especially on opposite sides of the signals. A more symmetrical multimodal integration pattern would be expected to include an increase in the proportional length between signals, and increased co-timing of multimodal signal onsets and offsets.

A third principle that is potentially relevant to multimodal integration patterns is the *principle of area*, which states that there is a tendency to group elements in a manner that results in the smallest visible figure, or the briefest temporal interval. More specifically, this Gestalt principle would predict a larger number of users who are simultaneous rather than sequential integrators,

or a larger overall ratio of simultaneous to sequential constructions. In any given multimodal integration pattern, Gestalt theory also would maintain that more than one principle can operate at the same time.

The single general principle underlying all Gestalt tendencies is the creation of a balanced and stable perceptual form that is capable of maintaining its *equilibrium*, just as the interplay of internal and external physical forces shape an oil drop [3, 4]. During multimodal communication, any factors that threaten the user's ability to achieve a goal, including threats to communication or task success, would be expected to set up a state of tension or disequilibrium. Under these circumstances, Gestalt theory would predict that an attempt would be made to fortify these basic organizational phenomena in order to restore balance [3, 4]. These basic Gestalt principles for organizing multiple elements are believed to contribute to *redundant coding* between the elements, which increases the speed and accuracy of perceptual processing.

Although these Gestalt views have constituted a theory of perception, the present research explores the potential relevance of these concepts to users' *production* of multimodal communication. In the context of interactive communication, it is well known that speakers often tailor the language they produce in a manner that accommodates a listener's perceptual capabilities. For example, the characteristics of hyperarticulate speech adaptations (*e.g.*, durational and articulatory) have been documented to enhance listeners' speech perception [13]. As a result, one might expect to find broader applicability of Gestalt principles to both the perception and production of interactive multimodal communication patterns.

1.3 Goals of the Present Research

The present study investigates the relevance of the Gestalt principles described above to the temporal synchrony of users' speech and pen multimodal integration patterns while interacting with a next-generation map system. Since previous research has documented that users are distributed bimodally into predominantly sequential versus simultaneous integrators, one specific goal was to assess the potential malleability of individual users' dominant integration patterns. To support this, a new dual-wizard simulation method was introduced, which was used to log a person's integration pattern in real time, to quickly identify their dominant integration pattern at the beginning of a session, and then to adapt the system's simulated error rate to encourage switching over to the non-dominant pattern. For example, if someone was predominantly a sequential integrator, then strong selective reinforcement was provided whenever they delivered a simultaneously-integrated multimodal construction (0% error rate) rather than their usual sequential pattern (40% error rate). Given this context, a Behavioral/Structuralist perspective would expect that users either would switch to using their non-dominant integration pattern, or at least increase the percentage of non-dominant constructions between the first and second half of a session. In contrast, the Gestalt principles outlined above would predict that users would apply consistent co-timing of their multimodal signals, including onsets and offsets, as an organizational cue to their relatedness. In fact, the high 40% system error rate should precipitate a state of *disequilibrium* in users, causing them to fortify or entrench further in their existing pattern of co-timing.

A second goal of this study was to model variation in users' multimodal integration patterns during episodes of system error handling. Past research on speech hyperarticulation has consistently reported that users' speech shifts to hyper-clear acoustic-prosodic features when they encounter system recognition errors and must repeat their input. One major feature of such hyperarticulation is an increase in total utterance duration, typically ranging 9-19% [11,13]. However, to date there has not been parallel work on hyper-clear communication patterns during unimodal pen or multimodal interaction. The present study was in part designed to examine how users would adapt a multimodal construction that incorporates speech as just one of multiple communication modalities. A Structuralist perspective that focuses on the analysis of a single modality would expect that elongation of the speech signal should replicate when speech is combined multimodally. In contrast, Gestalt theory would predict that hyperarticulation in a multimodal context could well take on a qualitatively different form, for example with adaptations in co-timing now becoming the main vehicle conveying change. In this case, elongation of the individual speech signal might be attenuated or absent altogether. Furthermore, Gestalt theory would predict that actual error handling episodes, like a high base-rate of system errors, should precipitate a state of disequilibrium, causing users to fortify their existing co-timed pattern.

A third specific goal of this research was to examine change in users' multimodal integration patterns while completing tasks varying in difficulty. In this study, participants interacted with map tasks that varied from low to very high difficulty with respect to the spatial intensiveness of information. Once again, Gestalt theory would predict that the demands generated by increasingly difficult tasks should precipitate a state of disequilibrium, which would cause users to progressively fortify their co-timed patterns.

The long-term goal of this research is the development of a coherent general theory of multimodal integration during human-computer interaction. This research also aims to contribute temporal models of multimodal integration patterns, which will play a critical role in establishing the temporal constraints needed for optimal multimodal signal fusion and for building a new generation of adaptive time-sensitive multimodal architectures.

2. METHODS

2.1 Subjects

There were 12 adult subjects, aged 21-58, five male and seven female. All were native speakers of English and paid volunteers. None of the subjects were computer scientists, and participants had varying degrees of computer experience.

2.2 Task

Subjects were presented with a scenario in which they acted as non-specialists working to coordinate emergency resources during a major flood in Portland, Oregon. To perform this task, they were given a multimodal map-based interface on which they received textual instructions from headquarters. They then used this interface to deliver multimodal input to the system by both speaking and making an appropriate mark on the map. Tasks involved gathering information (*e.g.,* "Find out how many sandbags are at Couch School Warehouse"), placing items (*e.g.,*

"Place a barge in the river southwest of OMSI"), creating routes (*e.g.,* "Make a jeep route to evacuate tourists from Ross Island Bridge"), closing roads (*e.g.,* "Close Highway 84") and navigating on the map (*e.g.,* "Move north on the map").

Figure 2 shows a screen shot of the interface used in the experiment. In this example, the message from headquarters was "Show the railroad along the east waterfront between Broadway Bridge and Fremont Bridge." Each command was designed for multimodal input. For example, a user working with the task in Figure 2 might say "This is the railroad" and draw a line along the river on the map (see Figure 2, Area b).

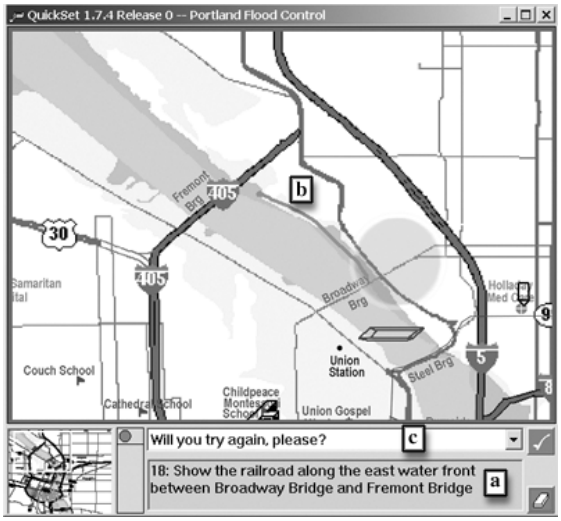

Figure 2. Flood management user interface

The tasks included four levels of difficulty: low, moderate, high and very high. Low difficulty tasks required the user to articulate just one piece of spatial-directional information (*e.g.*, north, west), or one location (*e.g.,* Cathedral School). Moderate difficulty tasks contained two pieces of spatial-directional/location information, high difficulty tasks contained three pieces, and very high difficulty tasks contained four such pieces. Table 1 lists sample commands from each difficulty level.

Table 1. Examples of task difficulty levels, with spatial-location lexical content in italics

Difficulty	Message from Headquarters
Low	Situate a volunteer area near ***Marquam Bridge***
Moderate	Send a barge from ***Morrison Bridge barge area*** to ***Burnside Bridge dock***
High	Draw a sandbag wall along ***east riverfront*** from ***OMSI*** to ***Morrison Bridge***
Very High	Place a maintenance shop near the ***intersection of I-405*** and ***Hwy 30*** just ***east*** of ***Good Samaritan***

2.3 Procedure

Volunteers were first oriented to the system and task by an experimenter, who provided instructions, answered questions, and offered feedback or help. Instruction was given until the subject was fully oriented and ready to work alone, which typically took about ten minutes. Following this orientation, the experimenter left the room and users began their session. The first ten tasks comprised an *identification band* that was used to determine the user's natural multimodal integration pattern (*i.e.*, their tendency to deliver the two input modes in a simultaneous or sequential manner). After this, users completed another 82 tasks.

Task instructions were delivered by the system in textual form on the lower part of the screen (see Figure 2, Area a), below a map of the appropriate area of Portland (Area b). There also was a text area for system feedback (Area c), in which confirmation or error messages were displayed. Users were told to complete tasks with this map-based system using their own words, and to use both pen and speech to communicate each command. The experimenter's instructions were unbiased with respect to how users could integrate modalities. If participants asked, they were told that they could use speech and pen input in any way they wished, as long as they used both modalities for each task. The system was introduced to users as an open-microphone implementation, so they did not need to tap the pen on the screen before speaking.

Upon completion, volunteers were interviewed about their interaction with the system, their integration pattern, and any errors they experienced, and were debriefed on the purpose of the study. Until that point, all users believed they were interacting with a fully functional system. The experiment lasted about one hour per participant.

2.4 Simulation Technique

2.4.1 Dual Wizard Technique

A novel simulation technique was developed for the present research that used two "wizards," or simulation assistants. The wizards, an Input Wizard and an Output Wizard, worked autonomously in separate rooms.

The Input Wizard's function was to make an initial determination of a user's natural integration pattern based on the ten-command identification band, and then to record the users' multimodal integration pattern in real time throughout the session. To do this, the Input Wizard observed the user's integration pattern by video feed during each task interaction, and recorded this pattern as either simultaneous or sequential. This information then was routed to both a data log and to the Output Wizard's system. Subsequent fine-grained video analyses determined that this real-time Input Wizard assessment was correct 99% of the time.

In a nearby room, the Output Wizard monitored the content of the user's input and then responded with appropriate feedback that was sent directly to the user's display. In addition to sending confirmations, the Output Wizard could send tailored feedback, such as "please use both pen and speech" or "use your own words," as needed. All Output Wizard messages were pre-scripted to expedite simulation response times. Finally, although the system required input from both wizards before it could proceed, no explicit coordination between the wizards was necessary.

2.4.2 Random Error Generator

The system included a random error generator that simulated system errors. When triggered, this mechanism occasionally overrode the system's response (transmitted by the Output Wizard), and instead responded with a failure-to-understand system message, such as "Will you try again, please?" Error

messages were delivered in the system feedback area (Figure 2, Area c), which was momentarily highlighted in red.

While subjects completed the initial ten-command identification band, the error generator delivered a fixed 20% error rate. Afterwards, the system paused briefly while the user's integration pattern was determined, which was based on a minimum of six out of ten commands exhibiting a simultaneous or sequential multimodal integration pattern. The random error generator then was set to deliver a 40% error rate, randomly distributed across commands, whenever the subject used their natural integration pattern. In contrast, a 0% error rate was delivered whenever the subject used their non-dominant pattern. For example, if a user delivered eight sequentially-integrated multimodal commands during their identification band, they were classified as a sequential integrator. Subsequently, they received a 40% error rate for any sequential commands, but no errors for simultaneous ones. Using this simulation software and contingent delivery of errors, the present study investigated whether users could be trained to switch from their natural multimodal integration pattern to the non-dominant pattern.

2.5 Research Design

As shown in Figure 3, the research design involved an initial pre-training phase during which a user's dominant multimodal integration pattern was identified during the first 10 commands, followed by two 41-command phases during which the user's integration pattern was tracked. These latter two phases were summarized separately for the first versus second half of the main session. The specific nature of training was contingent on a user's dominant integration pattern, as described in section 2.4.2. Based on this, the twelve subjects were divided into sequential integrators (SEQ), for whom training was designed to encourage simultaneous constructions, and simultaneous integrators (SIM), for whom training encouraged sequential integrations.

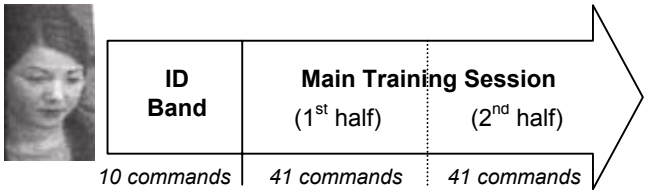

Figure 3. Research design summary

During the main training phases, two additional within-subject factors were evaluated: error handling and task difficulty. For error handling, a user's multimodal integration pattern was compared when a user originally input information (*i.e.*, no errors occurred), during first repetitions following a single system error, and during second and further repetitions when a system error persisted (*i.e.*, spiraled). For task difficulty, a user's multimodal integration patterns were analyzed across tasks varying in difficulty, including low, moderate, high, and very high tasks.

2.6 Data Capture and Transcript Coding

All sessions were videotaped, and multimodal speech and gesture data were analyzed for multimodal integration patterns and temporal synchronization, as well as for the degree of consistency and early predictability exhibited in users' patterns over their session. In addition, for each multimodal command, the start and end times of the pen and speech signals were hand-coded using SVHS video editing equipment to the nearest .1 second. Data from two previous studies have shown that inter-coder reliability for these measurements is accurate to the nearest .1 second [12, 14]. These time codes were collected from the video analysis and entered into a customized Access database interface. This flexible interface permitted checking the data for valid input and was designed to filter, compare, and summarize data. Based on the speech and pen time codes, this analysis tool also calculated the following durational information for each multimodal interaction:

Absolute Intermodal Overlap/Lag – During simultaneous integrations, the absolute duration of signal overlap was summarized in milliseconds (ms) for each multimodal command. During sequential integrations, the duration in ms from the end of the first signal to the start of the second one was summarized.

Intermodal Overlap Ratio – During simultaneous integrations, the ratio of signal overlap was calculated by dividing the absolute intermodal overlap by the total multimodal signal duration.

Intermodal Onset Differential – For each simultaneous multimodal command, the time in ms was recorded between the beginnings of the first and second input mode signals.

Intermodal Offset Differential – For each simultaneous multimodal command, the time in ms was recorded between the ends of the first and second input mode signals.

Speech Duration – For each multimodal command, the duration of the speech signal was recorded in ms. For the subset of verbatim matched speech analyses, only original input and first repetition pairs with identical lexical content were analyzed.

Pen Duration – For each multimodal command, the duration of the pen signal was recorded in ms. For the subset of verbatim matched analyses, only original input and first repetition pairs involving the same type of gesture or pen mark were analyzed (*e.g.*, two arrows).

Multimodal Command Duration – The total duration of each multimodal command was recorded in ms from the start of the first signal to the end of the final signal.

2.7 Reliability

Interlabeler agreement was calculated for independent coders. Measurements of start and end of the speech and pen signals were compared between the coders, and 89% of all measurements matched to within .1 second. When broken down by mode, 85% of pen measurements and 92% of speech measurements were reliable to within .1 second.

3. RESULTS

The following sections summarize users' multimodal integration patterns, as well as changes in the temporal synchronization of their integration patterns over the course of the session, during error handling, and as a function of task difficulty. A total of 1615 multimodal constructions were analyzed, or an average of 135 for each of the 12 subjects.

3.1 Multimodal Integration Patterns

One subject displayed a predominantly sequential integration pattern, while the other 11 subjects were simultaneous integrators. The average within-subject integration pattern consistency was 97% during both the initial identification band (range 83-100%) and also during the main task (range 81-100%). No subject

switched their dominant integration pattern to the non-dominant one over the course of the session, in spite of the high 40% error rate they received when they continued using their natural integration pattern. For sequentially-integrated commands, 83% involved pen input delivered before speech.

3.2 Temporal Synchronization Patterns

3.2.1 Absolute Intermodal Signal Overlap/Lag

For subjects who were simultaneous integrators, the *absolute signal overlap* increased from a mean of 1.41 seconds to 1.87 seconds between the first and second half of the session, or a relative increase of 33%. Correspondingly, for the sequential integrator the *absolute intermodal lag* increased from a mean of 0.27 seconds to 0.44 seconds, a relative increase of 63%, as shown in Table 2. An *a priori* paired t-test on logged absolute values confirmed that subjects *entrenched* significantly in whatever their dominant integration pattern was, with simultaneous integrators exhibiting increased signal overlap and sequential ones increased lag during the second half of their session, $t = 1.81$ (df = 11), $p < .05$, one-tailed.

During error episodes, absolute intermodal overlap/lag (logged values) entrenched significantly between a user's original input attempt and their first repetition following an error, *a priori* paired t-test, $t = 1.82$ (df = 11), $p < .05$, one-tailed. Likewise, the contrast between original input and second or subsequent repetitions was significant, *a priori* paired t-test on absolute logged values, $t = 3.28$ (df = 11), $p < .0035$, one-tailed. Table 2 summarizes that simultaneous integrators increased their signal overlap from 1.50 to 1.68 to 1.77 seconds during error spirals, a total relative increase of 18%, while the sequential integrator correspondingly increased their lag from .45 to .49 to .63.

Table 2. Average increasing signal overlap in seconds for SIM integrators and increasing lag for SEQ integrators

Integration Pattern	First Half	Second Half	Original Input	First Repeat	Deeper Repeats
SIM	1.41	1.87	1.50	1.68	1.77
SEQ	0.27	0.44	0.45	0.49	0.63

With respect to task difficulty, mean absolute signal overlap/lag revealed no significant change between the low and moderate difficulty levels, $t < 1$. However, a significant increase was observed between moderate and high task difficulty, *a priori* paired t-test on logged absolute values, $t = 3.02$ (df = 11), $p < .006$, one-tailed. An additional significant increase was observed between high and very high difficulty levels, $t = 1.81$ (df = 11), $p < .05$, one-tailed. Table 3 summarizes this progressive integration pattern entrenchment with increased task difficulty for all data.

Table 3. Average signal overlap in seconds for SIM integrators and lag for SEQ integrators with increased task difficulty

Integration Pattern	Low	Moderate	High	Very High
SIM	1.18	1.23	2.43	2.98
SEQ	.26	.42	.50	.33

An additional analysis compared a subset of this absolute signal overlap/lag data as a function of task difficulty, but in this case carefully matched the four task difficulty levels by type of task (*e.g.*, placing an object, creating a route). For this subset, mean absolute signal overlap/lag once again increased progressively with task difficulty, this time with a significant increase between low and moderately difficult tasks, *a priori* paired t-test, $t = 2.49$ (df = 35), $p < .009$, one-tailed. An additional significant increase was found between moderate and high difficulty tasks, $t = 2.50$ (df = 34), $p < .009$, one-tailed. An additional increase was observed between the high and very high levels, although not a significant one, $t = 1.10$, N.S. For this more carefully matched subset, the average signal overlap/lag increased from 1.10 to 1.75 seconds between low and very high task difficulty, or a total of 59%.

3.2.2 Intermodal Overlap Ratio

For subjects who were simultaneous integrators, the ratio of total speech and pen signal overlap to total multimodal signal duration increased from a mean of .50 in the first half of the session to .56 in the second half, a significant increase by *a priori* paired t-test, $t = 2.71$ (df=10), $p < .015$, one-tailed.

Table 4. Average percentage intermodal overlap for simultaneous integrators

First Half	Second Half	Original Input	First Repeat	Deeper Repeats
50%	56%	52%	54%	56%

During error handling, subjects' overlap ratios did not increase significantly between original attempts and first repetitions, $t = 1.50$, N.S. However, they did increase significantly between original attempts (.52) and deep spirals (.56) (*i.e.,* second repetitions and beyond), *a priori* paired t-test, $t = 2.17$, (df = 10), $p < .003$, one-tailed.

3.2.3 Intermodal Onset Differential

For subjects who were simultaneous integrators, the *intermodal onset differential* decreased significantly or the onsets moved closer together during error episodes, with the means changing from .83 to .66 to .59 seconds between original, first, and deeper repeats, respectively. This represented a significant decrease in signal onsets (log) between both original and first repeats, *a priori* paired t-test, $t = 4.63$, (df = 10), $p < .0005$, one-tailed, and between original and deeper repeats, $t = 2.48$, (df = 10), $p < .02$, one-tailed. The total relative decrease in onsets between original and deeper repeats was 29%. The first row of Table 5 summarizes these results. Finally, no significant change was found in onsets between the first and second half of a session, $t < 1$, N.S.

Table 5. Changes in mean onset & offset duration in seconds

	First Half	Second Half	Original Input	First Repeat	Deeper Repeats
Onset	N.S.		0.83	0.66	0.59
Offset	0.85	0.66	N.S.		

3.2.4 Intermodal Offset Differential

For subjects who were simultaneous integrators, the second row of Table 5 shows that the average *intermodal offset differential*s decreased from .85 to .66 seconds between the first and second half of a session, a relative decrease of 22%. An *a priori* paired t-test on offset differentials (log) confirmed this significant decrease, $t = 4.19$, (df = 10), $p < .001$, one-tailed. However,

during error episodes the offset differential did not change significantly, $t < 1$, N.S.

3.2.5 Speech Duration

Speech duration did not change significantly from the first to second half of a session, $t = 1.38$, N.S., nor did it change significantly when all data were compared during error handling, $t < 1$, N.S. However, when a subset of 115 speech durations were compared that had been matched on *verbatim* lexical content, original versus first repeats averaged 1.65 and 1.73 seconds, a 4.8% relative increase, which was significant by *a priori* paired t-test, $t = 2.37$, (df = 114), $p < .01$, one-tailed.

3.2.6 Pen Duration

Pen duration likewise did not change significantly between the first and second half of a session, $t < 1$, N.S. During error handling, no significant change in pen duration was found between original input to first repeats, $t < 1$, N.S. However, pen duration (log) did increase from 1.76 to 1.92 seconds between original input and deeper repeats by *a priori* paired t-test, $t = 2.93$, (df = 11), $p < .01$, one-tailed. This represented a 9% relative increase in pen duration during deeper spiral errors. However, on a subset of pen durations that contained carefully matched gestural marks, no significant change was found in duration (log) between original input and first repetitions, $t < 1$, N.S.

3.2.7 Multimodal Command Duration

There were no significant changes in total multimodal signal duration between the first and second half of a session, $t = 1.37$, N.S., nor were there any significant changes during error handling, $t < 1$, N.S.

3.3 Post-Experimental Interview

Based on post-experimental interview data, 67% of participants were able to describe their basic multimodal integration pattern, whereas the remaining 33% were completely unaware of it. The majority of participants, or 83%, spontaneously commented about the system errors that they experienced. Most believed that systems errors were caused by their spoken word choice (58%), type of pen marks (42%), speech articulation (33%), or multimodal signal co-timing (33%). For example, one user reported about their co-timing, "…when I drew a line I would draw the line the whole time I was saying the sentence and end at the exact same time." A few users also mentioned that they believed consistency was important in their interaction style with the computer.

4. DISCUSSION

This research summarizes a comprehensive body of evidence demonstrating that changes in *co-timing* provide the main organizational cues for binding speech and pen input during multimodal communication. Furthermore, users' dominant multimodal integration pattern was strikingly consistent (97%) and resistant to change, even when strong selective reinforcement was delivered to encourage switching from a sequential to simultaneous integration pattern, or vice versa. Instead, both sequential and simultaneous integrators showed evidence of *entrenching* further in their dominant integration patterns (*i.e.*, increasing their intermodal *lag* during sequential integrations and *overlap* during simultaneous integrations) over the course of an interactive session, during system error handling, and when completing increasingly difficult tasks. As shown in Table 2 (left side), subjects who were simultaneous integrators increased their signal *overlap* significantly from 1.41 to 1.87 seconds between the first and second half of the session, a relative increase of 33%. Likewise, for sequential integrators the *absolute intermodal lag* increased from .27 to .44 seconds, or by 63%. Convergent results based on an overlap ratio further underscored these findings. The high degree of consistency in multimodal integration patterns may in part have reflected users' adoption of a "success strategy" in interactions with the computer. However, these systematic changes in multimodal signal co-timing occurred in spite of the fact that users varied widely in explicit awareness of their own integration patterns.

In the context of error handling, this entrenchment of multimodal signal co-timing was the dominant feature of users' hyper-clear multimodal language, creating a qualitatively new phenomenon of *multimodal hypertiming*, unlike the unimodal speech hyperarticulation reported previously during human-computer interaction. Table 2 (right side) shows the pattern of co-timing entrenchment that occurred specifically during error handling episodes, with the signal overlap for simultaneous integrators increasing significantly from 1.50 to 1.77 seconds between original input and second or later repetitions, or by 18%. Similarly, sequential integrators increased their lag from .45 to .63 seconds, or 40%. Once again, convergent findings emerged from the data on intermodal overlap ratios. In the present data, these unique multimodal temporal changes now overshadowed the attenuated speech elongation previously reported for unimodal speech hyperarticulation. In previous reports, the total utterance duration for spoken language has consistently shown relative increases of 9-19% when the same person repeats the same lexical content after a system error [11, 13]. However, during the present multimodal communications total speech utterance duration on matched verbatim lexical content increased less than 5%, and total pen duration did not increase at all.

Users' multimodal synchronization patterns also showed a clear pattern of entrenchment as the present map-based spatial tasks became progressively more elevated in difficulty. In fact, multimodal signal overlap/lag increased steadily from 1.10 to 1.75 seconds for tasks varying from low to very high difficulty, or by 59% total. To summarize, as spatial tasks become more challenging, both sequential intermodal lags and simultaneous overlaps increased.

From a theoretical viewpoint, Gestalt principles correctly predicted that users would apply consistent co-timing of their multimodal signals as an organizational cue to their relatedness. Clearly, the temporal dimension of multimodal organization becomes especially important for establishing the relatedness of multimodal signals when one or more of the modalities involved has no spatial instantiation. In fact, from a Gestalt perspective the high 40% error rate would have precipitated a state of *disequilibrium* in users, causing them to fortify or entrench further in their existing co-timing patterns. This included an increase in the symmetrical co-timing of signal onsets and offsets under different circumstances, as observed in Table 5. In addition, when users' communication was specifically threatened by system errors or by increasingly demanding spatial tasks, Gestalt theory correctly predicted that they would fortify these aspects of their multimodal signal co-timing in order to restore the equilibrium

needed to support a listener's successful perceptual processing. Finally, Gestalt theory predicted that simultaneous users and integration patterns would be more prevalent than sequential ones, due to the bias toward creation of briefer rather than longer temporal intervals. As shown in Figure 1, 70% of users from children through the elderly are predominantly simultaneous integrators. In summary, the Gestalt principles of *proximity, symmetry*, and *area,* as well as the general concept of *disequilibrium*, all have provided a valuable framework for understanding the present speech and pen multimodal integration patterns. In this research, the theoretical utility of these principles has been demonstrated for the *production* of multimodal communication patterns, rather than the perceptual focus of most previous Gestalt research.

One implication of these results is that future computational systems will need to accurately model users' existing multimodal integration patterns, including the major parameters like system errors and task difficulty that cause these patterns to vary systematically, rather than naively assuming that users can be trained to adopt to a particular style. The basic data on users' multimodal integration patterns present very fertile opportunities for adaptive processing, since users are divided into two basic types, with early predictability and very high consistency. Quantitative empirical modeling of the type described in this paper is expected to provide a scientific foundation for accurately predicting the major variation in users' integration patterns. Such models can be used to guide the development of new strategies for adapting temporal thresholds in future time-sensitive multimodal architectures during the fusion process, potentially yielding substantial improvements in system response speed, robustness, and overall usability.

The long-term goal of this research is the development of a coherent theory of multimodal integration during human-computer interaction. To achieve this goal, future research could benefit by exploring other specific implications of Gestalt theory for the organization of multimodal communication patterns. In addition, further research should begin exploring the generality of the present results on pen/voice multimodal integration patterns for other modality combinations.

5. ACKNOWLEDGEMENTS

Thanks to Kristy Hollingshead for assistance with data collection, scoring, and programming. Thanks also to Jim Ann Carter for expert graphics, and to members of CHCC for many insightful discussions. This research was supported by DARPA and NSF Grant No. IIS-0117868.

6. REFERENCES

[1] Benoit, J., C. Martin, C. Pelachaud, L. Schomaker & B. Suhm. Audio-visual and multimodal speech-based systems. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation* (D. Gibbon, I. Mertins & R. Moore, eds.), Kluwer, Boston MA, 2000, 102-203.

[2] Bregman, A.S. *Auditory Scene Analysis*. MIT Press, Cambridge MA, 1990.

[3] Koffka, K. *Principles of Gestalt Psychology*. Harcourt, Brace & Company, NY, 1935.

[4] Kohler, W. *Dynamics in Psychology*. Liveright, NY, 1929.

[5] Massaro, D. & D. Stork. Sensory integration and speech reading by humans and machines. *Amer. Scien.*, 1998, 86, 236-244.

[6] McGrath, M. & Q. Summerfield. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *JASA*, 1985, 77(2), 678-685.

[7] McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. Univ. of Chicago Press, Chicago IL, 1992.

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

[9] Oviatt, S.L. Ten myths of multimodal interaction. *CACM*, 1999, 42(11), 74-81.

[10] Oviatt, S.L. Multimodal Interfaces. *Handbook of Human-Computer Interaction* (J. Jacko & A. Sears, eds.), Law. Erlb., Mahwah NJ, 2003, 286-304.

[11] Oviatt, S.L., R. Coulston & C. Darves. Predicting children's hyperarticulate speech during human-computer error resolution. *Conf. of ASA,* Nashville TN., April 2003.

[12] Oviatt, S.L., A. DeAngeli & K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction, In *Proc. of CHI '97*, 415-422.

[13] Oviatt, S.L., G. Levow, E. Moreton & M. MacEachern. Modeling global and focal hyperarticulation during human-computer error resolution. *JASA*, 1998, 104(5), 1-19.

[14] Xiao, B., C. Girand & S.L. Oviatt. Multimodal integration patterns in children. In *Proc. of ICSLP'2002*, 629-632.

[15] Xiao, B., R. Lunsford, R. Coulston, M. Wesson & S.L. Oviatt. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences, to be presented at the *Fifth International Conference on Multimodal Interfaces*, Vancouver, B.C., Nov. 2003.