Modeling Multimodal Integration Patterns and Performance in Seniors: Toward Adaptive Processing of Individual Differences

Benfang Xiao Rebecca Lunsford Rachel Coulston Matt Wesson Sharon Oviatt
Oregon Health and Science University
OGI School of Science & Eng.
20000 NW Walker Road
Beaverton, OR 97006, USA
+1 503 748 7397

{benfangx, rebeccal, rachel, wesson, oviatt}@cse.ogi.edu

ABSTRACT

Multimodal interfaces are designed with a focus on flexibility, although very few currently are capable of adapting to major sources of user, task, or environmental variation. The development of adaptive multimodal processing techniques will require empirical guidance from quantitative modeling on key aspects of individual differences, especially as users engage in different types of tasks in different usage contexts. In the present study, data were collected from fifteen 66- to 86-year-old healthy seniors as they interacted with a map-based flood management system using multimodal speech and pen input. A comprehensive analysis of multimodal integration patterns revealed that seniors were classifiable as either simultaneous or sequential integrators, like children and adults. Seniors also demonstrated early predictability and a high degree of consistency in their dominant integration pattern. However, greater individual differences in multimodal integration generally were evident in this population. Perhaps surprisingly, during sequential constructions seniors' intermodal lags were no longer in average and maximum duration than those of younger adults, although both of these groups had longer maximum lags than children. However, an analysis of seniors' performance did reveal lengthy latencies before initiating a task, and high rates of self talk and task-critical errors while completing spatial tasks. All of these behaviors were magnified as the task difficulty level increased. Results of this research have implications for the design of adaptive processing strategies appropriate for seniors' applications, especially for the development of temporal thresholds used during multimodal fusion. The long-term goal of this research is the design of high-performance multimodal systems that adapt to a full spectrum of diverse users, supporting tailored and robust future systems.

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

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: 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, senior users, task difficulty, human performance errors, self-regulatory language

1. INTRODUCTION

One important theme of contemporary computing is improving the flexibility and appropriate tailoring of systems for diverse user groups, tasks and environments. Multimodal interfaces [15], multi-agent systems [6], and speech recognition systems that are capable of speaker and environmental adaptation [7], all are examples of efforts aimed at improving both the reliability and usability of systems for real-world usage contexts. In the case of multimodal interfaces, they can be designed to support simultaneous use of input modes, or to permit switching among modes in order to take advantage of the modality best suited to a particular task, environment, or set of user preferences and capabilities [15]. Compared with traditional keyboard and mouse interfaces, multimodal interfaces also permit richer and more varied user expressiveness via relatively natural input modes. Of course, this incorporation of natural input modes like speech, gaze, and pen-based writing and gesturing also entails greater individual differences between users in their natural patterns of self expression, which means that the development of adaptive multimodal interfaces will be a particularly important area for future research.

At the present time, adaptive processing of multimodal interfaces is still in its infancy. Examples of recent work on this topic include adaptively estimating stream weights at different noise levels in audio-visual speech recognition [5], and adaptively delivering multimodal feedback based on senior users' visual abilities [9]. However, many other important aspects of adaptive processing remain to be considered in the design of new multimodal interfaces. For example, in the case of elderly users, multimodal interfaces will need to adapt to the large individual

differences known to be present in this population, as well as to performance limitations associated with slower processing speeds and memory decline [4, 17].

One especially central issue for all of multimodal interface design will be accurate modeling and adaptation of multimodal interfaces to users' multimodal integration patterns. Present state-of-the-art systems like QuickSet [2] have been developed to accommodate typical adults' pen/voice multimodal integration patterns [16]. Such systems currently use *fixed temporal thresholds* to determine when modality fusion is "legal" [15]. These temporal constraints play an essential role in resolving when parallel or sequentially-delivered speech and pen signals should be integrated into a whole multimodal interpretation of user input. However, multimodal systems based on fixed temporal thresholds, especially if they rely on typical adult data, may be inaccurate for other user groups like children and seniors. Fixed temporal thresholds also do not permit any tailoring or optimization that may be needed for handling individual differences within user groups, or differences among tasks and usage environments. In response to these limitations, one key direction for future multimodal interface design will be the development of a new class of *adaptive temporal thresholds*, which will require data collection and lifespan modeling of individuals' natural modality integration patterns in different tasks**.**

Previous Literature on Multimodal Integration Patterns
In previous research on children [20] and adults [16] involving pen/voice multimodal interaction, individuals were observed to integrate their pen and speech input either *simultaneously* (i.e., with speech and pen signals overlapped) or *sequentially* (i.e., signals separated by a brief lag). As shown in Table 1, past research has documented that children have a predilection to integrate speech and pen simultaneously (77%), whereas adults tend to integrate sequentially (64%). Furthermore, all children and adults have demonstrated a dominant integration, which was predictable for 92% of children and 100% of adults on their very first multimodal construction [13, 20]. Users' dominant integration patterns were deployed with very high consistency, or for 93.5% of children's multimodal constructions and 90% of adults' constructions during system interactions.

During sequentially-integrated multimodal constructions, the previous literature has indicated that pen input precedes speech 97-99% of the time for children and adults, respectively. Children's and adults' average intermodal lag during pen/voice sequential integrations also is the same, or 1.1 seconds (secs). However, as shown in Figures 1 and 2, adults' sequential lags range twice as long as children's, or 4.1 rather than 2.1 secs.

Previous Literature on Aging and Performance
Age-related decline in performance speed has been well documented in previous research [4, 10, 17]. For example, reaction times and response latencies generally have been observed to increase with age, and these effects are well known to interact with increasing task difficulty in seniors [4, 10, 17]. Increased latencies most frequently are observed in seniors for tasks requiring decision making, problem solving, language comprehension, and other forms of more complex reasoning [10, 17, 19]. Whereas reading rates per se are not necessarily slower in seniors, nonetheless total reading time and the likelihood of making comprehension errors do increase, especially as the complexity of a text increases [19].

Table 1. Percentage of simultaneously-integrated multimodal constructions (SIM) versus sequentially-integrated ones (SEQ) for individual children, compared with adults

Children			Adults		
User	**SIM**	**SEQ**	**User**	**SIM**	**SEQ**
SIM integrators:			**SIM integrators:**		
1	100	0	1	100	0
2	100	0	2	94	6
3	100	0	3	92	8
4	100	0	4	86	14
5	100	0	**SEQ integrators:**		
6	100	0	5	31	69
7	98	2	6	25	75
8	96	4	7	17	83
9	82	18	8	11	89
10	65	35	9	0	100
SEQ integrators:			10	0	100
11	15	85	11	0	100
12	9	91			
13	2	98			
Average Consistency - 93.5%			**Average Consistency - 90%**		

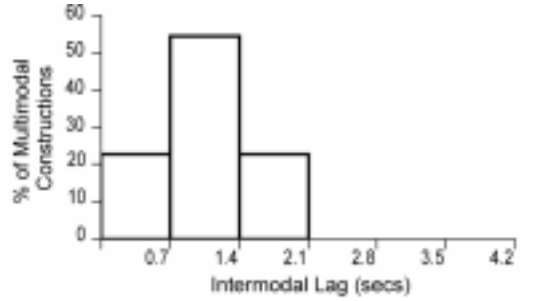

Figure 1. Distribution of sequential intermodal lags for children

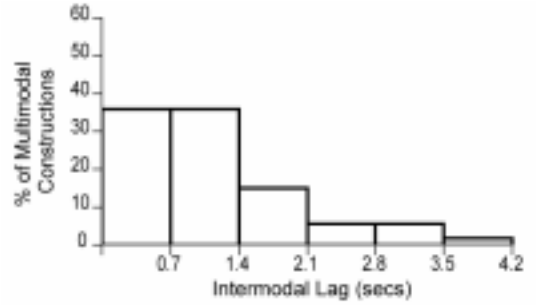

Figure 2. Distribution of sequential intermodal lags for adults

Extensive research has shown that short-term or working memory capacity also declines in adults after age 45, and this change has an impact on task performance across a wide variety of domains [18, 19]. Reductions in working memory capacity have been correlated with decreased performance on language comprehension tasks [11]. Essentially, seniors lose the ability to easily retain partial bits of information in working memory as they build up a complete semantic model of read information, and this loss of information through memory decay or displacement disrupts their comprehension [11, 17]. This interaction between memory and reading comprehension is especially problematic as the complexity of reading material increases [19]. In addition, reasoning and memory for spatial information, such as that

required in map-based tasks, is well known to be difficult and to decrease with advancing age [17]. Compared with younger adults, seniors typically perform both more slowly and less accurately on spatial tasks, and their accuracy on such tasks decreases with increased spatial complexity [17].

Given these speed and memory declines, seniors would be expected to require more time and to make more errors on map-based spatial tasks. Under circumstances that are similarly intellectually demanding, children have been documented to engage in a high rate of audible *self-regulatory language* [1, 12], although this behavior tends to be internalized during adulthood [12]. Basically, self-regulatory language, also known as *self talk* or *private speech*, is believed to be a kind of think-aloud process in which individuals verbalize poorly understood aspects of difficult tasks to assist in guiding their own cognition and actions [1]. As task difficulty increases, self talk serves as a scaffolding behavior that can support human performance. In fact, when individuals with Down's Syndrome were trained to use self talk while performing a memorization task, their memory spans increased significantly [3]. The presence of self-regulatory language is not only performance enhancing, but also characteristic of behavior throughout the lifespan whenever an individual is faced with a challenging task:

> "…the need to engage in private speech never disappears. Whenever we encounter unfamiliar or demanding activities in our lives, private speech resurfaces. It is a tool that helps us overcome obstacles and acquire new skills." (*L.E. Berk, 1994, pp 80*).

Goals of current study

One goal of the current research was to analyze the multimodal integration patterns of seniors during map-based tasks, and to compare their integration patterns with those reported previously for children and younger adults. Such data are needed to create accurate temporal models for use in multimodal systems that are expected to support cross-generational software applications, and also to begin designing effective adaptive strategies for a new class of advanced multimodal systems. With respect to integration patterns, it was predicted that seniors would:

- Demonstrate either a simultaneous or sequential dominant integration pattern,
- Display their dominant integration pattern very early during a system interaction,
- Remain highly consistent in their dominant pattern throughout a system interaction.

In light of the literature on performance and aging, it also was predicted that seniors would:

- Tend to integrate modes in a slower sequential manner, compared with children and adults,
- Exhibit longer intermodal lags than children and adults,
- Show increases in their intermodal lags as task difficulty becomes elevated.

Another goal of the present research was to analyze aspects of seniors' performance, as well as the relation of their performance to task difficulty levels. In particular, task-critical human performance errors, task response latencies (i.e., time required to initiate a task after instructions arrive), and self talk (i.e., speech verbalized before multimodal input to the system is delivered) all were examined in the present study. Given the combined effects of decreased motor and processing speed, diminished memory capacity, and challenging spatially-oriented tasks, it was expected that seniors would require more time to plan and initiate their tasks, would exhibit relatively high rates of task-critical performance errors, and also would produce elevated rates of self-regulatory speech as they attempted to process and retain spatial information during map tasks. Finally, all of these performance measures were predicted to increase in seniors as the spatial complexity of map tasks increased.

2. METHODS

2.1 Subjects

There were fifteen senior subjects aged 66 to 86 years, six male and nine female. All were native speakers of English and paid volunteers. None of the subjects were computer scientists, and they had varying degrees of computer experience from none to basic E-mail and office processing skills. All subjects were healthy, without any major cognitive deficits, physical limitations, or chronic diseases. All seniors also were living independently, and were physically active within the local community. The educational background of the subjects ranged from high school graduates to Bachelor's degrees. They also were from diverse professional backgrounds, such as nursing, property management, and real estate.

2.2 Scenario

Subjects were presented with a scenario in which they were to act as non-specialists coordinating emergency resources during a major flood in Portland, Oregon. They were given a multimodal map-based interface on which they received textual instructions from headquarters. They then used this interface to deliver instructions to the map system using both speech and pen input. Individual tasks involved obtaining information (e.g., "Find out how many sandbags are at Couch School Warehouse"), placing items on the map (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 controlling the map display (e.g., "Move north on the map").

Figure 3 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 water front between Broadway Bridge and Fremont Bridge." Each task was designed for multimodal input. For example, a subject working with the task in Figure 3 might say "This is the railroad" and draw a line along the river on the map (see Figure 3, Area b).

The tasks included three levels of difficulty: low, moderate, and high. Low difficulty tasks required the subject to articulate just one piece of spatial-directional information (e.g., north, west), or one location (e.g., Cathedral School). Each additional direction or location translated into one level of difficulty higher. Therefore, moderate difficulty tasks contained two pieces of spatial-directional/location information, and high difficulty tasks

contained three pieces. Table 2 shows sample tasks from each of these task difficulty levels.

Figure 3. User interface displaying a map-based flood management task

Table 2. Examples of task difficulty levels, with spatial-directional/location lexical content in italics

Task Difficulty	Instruction 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*

2.3 Procedure

Instructional prompts that described tasks were delivered as text instructions on the lower part of the computer screen (see Figure 3, area a), which was displayed below a map showing the related area of Portland (area b). There was also a text area for system feedback (area c), where confirmation or error messages were displayed. The subjects were told to tap the computer screen to engage the microphone before communicating a task, to express themselves naturally using their own words, and to use both pen and speech to communicate each task to the map system. Subjects were told that they could integrate speech and pen input in any way they wished when delivering their multimodal commands to the system, as long as they used both modalities for each task.

The subjects were first given training until they were fully oriented and ready to work alone. Typically, the training took about 15 minutes. However, four subjects required two training sessions, which lengthened their training to 20-35 minutes. Senior adults frequently require longer training times and more help with computer tasks than younger adults [4]. During the training session, an experimenter was present to give instructions, answer questions, and offer feedback and help. Following training, the experimenter left the room and the subjects completed their session independently, which involved 80 tasks.

Upon completion, the subjects were interviewed about their interaction with the system, any errors they experienced, and were debriefed on the purpose of the study. Until that point, all subjects believed they were interacting with a fully-functional computer system. The entire experiment lasted about an hour per participant, although one subject required 1 hour and 40 minutes.

2.4 Simulation Technique

The data collection process was based on a high-fidelity semi-automatic simulation technique similar to that used for previous studies involving adults [14] and children [20]. In the current simulation environment, the random error generator delivered a 5% task error rate.

2.5 Research Design, Data Capture and Coding

The experimental design involved 15 seniors, and within-subject data collection on tasks involving three levels of task difficulty: low, moderate, and high. All sessions were videotaped. Both temporal integration and performance measures were scored using SVHS video editing equipment. The following is a description of the scoring conducted for each dependent measure.

2.5.1 Temporal Synchronization Measures

Multimodal Integration Pattern

The integration pattern for each subject's first complete multimodal construction was scored for every task. When speech and pen signals overlapped, the construction was scored as simultaneous, and when no signal overlap was present the construction was scored as sequential. Subjects were classified as simultaneous or sequential integrators when 65% or more of their constructions predominantly fit one integration pattern. Others were classified as non-dominant.

Sequential Intermodal Lags

For sequential constructions, the intermodal lag was measured as the interval in milliseconds from the end of the first input mode to the start of the second mode. These measures were analyzed to determine average and maximum lag durations. The precedence relation between modes also was scored (i.e., whether pen preceded speech or speech preceded pen in order of delivery).

2.5.2 Performance Measures

The following performance measures also were assessed on a subset of 27 tasks that were matched on task difficulty level (9 low, 9 moderate, and 9 high difficulty). Human performance measures were analyzed to determine whether task difficulty had an impact on task-critical performance errors, response latency to plan and execute a task, and self talk during the preparatory phase before the user completed a task using multimodal input to the system.

Human Performance Errors

Task-critical human performance errors (HPEs) were scored whenever the subject specified an incorrect location, direction, or name for a location when completing their task, or if the task content was completely in error. HPEs were coded in part to validate that the task difficulty levels were in fact experienced by subjects as increasing in difficulty from low to moderate to high as subjects were required to keep track of additional spatial information. All tasks during a subject's session were coded for the total number of such errors, which then was converted to a

percentage of errors on tasks within the different task difficulty levels.

Self Talk
Self talk was scored whenever the subject talked or subvocalized audibly before actually tapping on the computer screen to enter their input. Each individual task was scored for the presence or absence of self task, which then was converted to the total percentage of tasks containing self talk within each task difficulty level.

Response Latency
Response latencies were measured as the duration between when a task instruction first appeared on the computer screen until the subject's first multimodal signal started when they entered their input to the computer during that task.

2.5.3 Reliability
Each multimodal construction was independently scored by a second scorer, who checked multimodal integration patterns, and intermodal lags. Inter-coder reliability on 80% of the intermodal lags was accurate to within .07 secs. This reliability also is comparable to that reported in previous studies on adults and children, for which intermodal lags matched to within .1 secs [16, 20].

Second scoring on approximately 15-20% of the performance data also indicated that 93% of scored HPEs and self talk matched perfectly. Finally, 80% of the task response latency durations matched to within .13 secs.

3. RESULTS
3.1 Temporal Synchronization Measures
In total, data on over 1150 multimodal constructions were available to be scored for the integration patterns described below, including data on approximately 200 sequential constructions containing intermodal lags. The average utterance length for users' multimodal constructions in this domain was 11 words.

3.1.1 Multimodal Integration Pattern
As shown in Table 3, 12 (92%) of the 13 seniors with a dominant pattern were simultaneous integrators, 1 (8%) was a sequential integrator, and the remaining 2 had no dominant pattern.

3.1.2 Consistency of Integration Pattern
Seniors consistently used their main integration pattern an average of 88.5% of the time during a session when delivering multimodal constructions. Of the 13 seniors who had a dominant integration pattern, 93.5% were consistent in their use of this pattern. Table 3 summarizes individual differences in integration pattern consistency.

3.1.3 Predictability of Integration Pattern
Of the 13 seniors who demonstrated a dominant pattern, all of them adopted this pattern within 2 out of the first 3 multimodal constructions. In fact, 85% were predictable on their very first multimodal input.

Table 3. Percentage of simultaneously-integrated multimodal constructions (SIM) versus sequentially-integrated constructions (SEQ) for seniors

Senior Users		
User	**SIM**	**SEQ**
SIM integrators:		
1	100	0
2	100	0
3	100	0
4	97	3
5	96	4
6	95	5
7	95	5
8	92	8
9	91	9
10	90	10
11	89	11
12	73	27
SEQ integrators		
13	1	99
Non-dominant integrators:		
14	59	41
15	48	52
Average Consistency – 88.5%		

3.1.4 Temporal Precedence of Input Modes
During sequentially-integrated multimodal commands, pen input preceded speech 64% of the time.

3.1.5 Intermodal Lags
Seniors' intermodal lags during sequentially-integrated multimodal constructions averaged 1.02 secs. As shown in Figure 4, 42% of all lags ranged between 0.0 and 0.7 secs, 72% between 0.0 and 1.4 secs, 90% between 0.0 and 2.1 secs, 96% between 0.0 and 2.8 secs, 98% between 0.0 and 3.5 secs, and 100% between 0.0 and 3.8 secs.

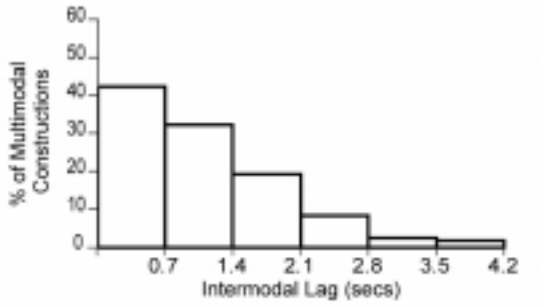

Figure 4. Distribution of intermodal lags for seniors in secs during sequential constructions

The average intermodal lag was .77 secs for seniors who were habitually simultaneous integrators, compared with 1.24 secs for those who were habitually sequential integrators. These senior data were compared with adult intermodal lags for habitually simultaneous versus sequential integrators, which were .78 secs versus 1.28 secs, as shown in Table 4 (S. Oviatt, personal communication, 2003). Independent t-tests confirmed that habitually sequential integrators had significant longer lags than simultaneous integrators for both adults, $t=1.71$, $(df=65)$ $p<.05$, one-tailed, and for seniors, $t=4.07$ $(df=189)$ $p<.0005$, one-tailed. The relative percentage increase in average lag between simultaneous and sequential integrators ranged 61-64% for these two groups, as shown in Table 4.

Table 4. Mean intermodal lags in secs for habitually sequential versus simultaneous integrators

User Population	SIM Mean Lag	SEQ Mean Lag	Relative Increase
Adult	.78	1.28	+64%
Senior	.77	1.24	+61%

Senior intermodal lags also increased significantly when task difficulty increased from low to moderate to high. As shown in Figure 5, mean lag durations were .99 secs during low difficulty tasks, but increased to 1.46 secs during moderate to high difficult tasks, which was a significant increase by independent t-test, t=2.07 (df=60) p<.025, one-tailed.

Figure 5. Impact of increasing task difficulty on average intermodal lags during sequential constructions

3.2 Performance Measures

Data on over 400 multimodal constructions were available to be scored for the human performance measures outlined below.

3.2.1 Human Performance Errors

During the course of their session, 87% of the subjects had task-critical performance errors. Human performance errors showed a significant increase between low and moderately difficult tasks, with 4.4% of the low difficulty tasks containing errors versus 10.3% of the moderately difficult tasks, *a priori* paired t-test, t=2.48 (df=14) p<.0135, one-tailed. There also was a significant increase in errors between the moderate versus high difficulty tasks, which averaged 10.3% versus 25.7%, *a priori* paired t-test, t=2.66 (df=14) p<.009, one-tailed. Figure 6 summarizes this increase in errors with task difficulty level.

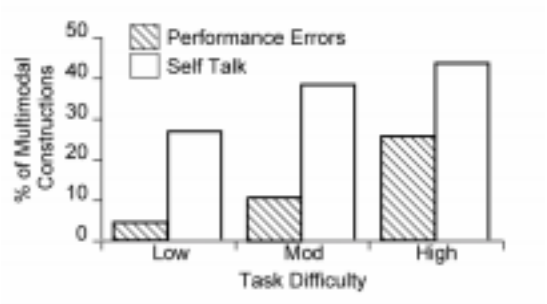

Figure 6. Impact of increasing task difficulty on percentage of task-critical performance errors and self talk

3.2.2 Self Talk

Over 80% of the subjects engaged in self talk at some point during their session. As shown in Figure 6, self talk increased significantly from low to moderate difficulty tasks, (26.9% versus 38.5%, respectively), *a priori* paired t-test, t=3.07 (df=14) p<.004, one-tailed. Self talk also increased between the moderate to high difficulty tasks (38.5% versus 43.7%), but not significantly so, paired t-test, t=1.52 (df=14), N.S.

In addition, there was a strong correlation between human performance errors and self talk, .66, which was significant, F=10.22 (df=1,13) p<.0035, one-tailed. In fact, 44% of the variance in self talk could be accounted for by knowing an individual senior's level of human performance errors, p^2_{xy}=0.44, (N=15). Figure 7 shows the best-fitting polynomial regression, with self talk increasing above 10% HPEs, and with a steep rise above approximately 18% HPEs.

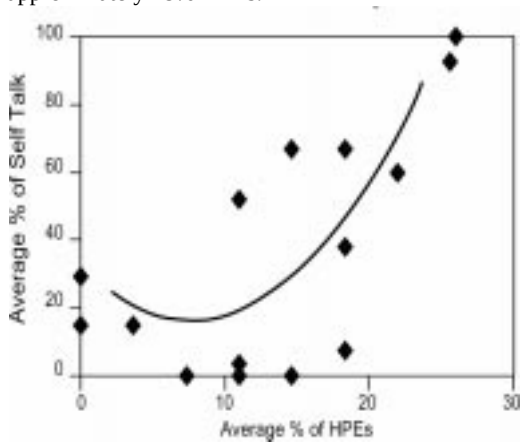

Figure 7. Linear regression between individual seniors' percentage of performance errors (HPEs) and self talk

3.2.3 Response Latency

Response latency averaged 14.3 secs, and ranged between 2.9 and 44.1 secs. As shown in Figure 8, response latency increased from an average of 9.4 secs during the low difficulty tasks, to 16.5 secs in the moderately difficult ones, and 18.3 secs during high task difficulty. This difference was significant between low and moderate difficulty levels, *a priori* paired t-test, t=8.31 (df=13) p<.0001, one-tailed, with no further significant increase between moderate and high levels, t=1.45, N.S.

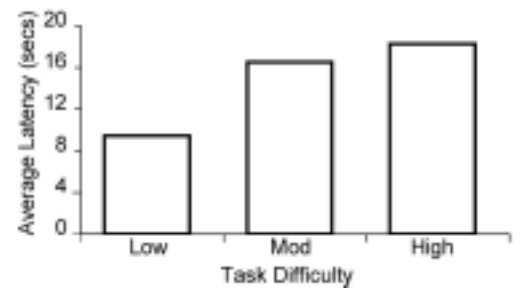

Figure 8. Impact of increasing task difficulty on average response latency

4. DISCUSSION

Pen/voice multimodal integration patterns for seniors replicated those observed previously in children and adults in several important respects. Most seniors, or 87%, could be classified as either predominantly simultaneous or sequential integrators. For those with a dominant integration pattern, their average consistency in using it was 93.5%, which is similar to 93.5% for children and 90% for adults. Additionally, for most seniors, or 85%, their dominant pattern was predictable on the very first multimodal input to a system, which also is similar to 92% for children and 100% for adults.

However, seniors as a group exhibited even greater individual differences than either children or adults. For example, whereas all children and adults in previous studies were clearly classifiable as having a dominant integration pattern, 13% of seniors in this study (2 out of 15) were not. As a result, when all seniors were included in the consistency data, their overall consistency as a group dropped to 88.5%, and the early predictability of their integration pattern based on first input dropped to 73%. Finally, although pen input preceded speech 97-99% of the time during sequential constructions for children and adults, for seniors the precedence of pen input was a far less consistent 64%. These individual differences among seniors indicate that both adaptable and adaptive processing strategies could provide a particularly fertile direction for the development of multimodal interfaces for "aging-in-place" and other senior applications.

Based on previous literature indicating that reaction and processing time both increase with advancing age, it was anticipated that average and maximum intermodal lags for seniors would be longer than either children or adults. However, seniors' intermodal lags averaged 1.02 secs, which was not slower than the 1.1 secs typical of both children and adults. In addition, the maximum intermodal lag observed in seniors was 3.8 secs, which did not exceed the 4.1 second maximum lag reported previously for younger adults. Likewise, when the intermodal lags from habitually simultaneous versus sequential senior integrators were compared with those of younger adults, they were virtually identical, as shown in Table 4. Finally, seniors as a group were not more likely to display a slower sequential integration pattern than either children or adults. In fact, 80% of seniors were simultaneous integrators, which was surprisingly similar to the 77% reported previously for children. One possible explanation for seniors' rapid intermodal lags during pen/voice multimodal interaction is that they reflect a *highly automated and reactive behavioral pattern,* which is learned like typing or driving a car, and therefore not subject to the same processing delays involved in tasks requiring decision making, memory, and other forms of cognitive processing.

In many respects, these data on individual differences in seniors' multimodal integration patterns present an ideal opportunity for adaptive processing. That is, like children and adults, seniors are divided into two basic types, with early predictability and relatively high consistency in their integration pattern. Furthermore, even when simultaneous integrators do display occasional sequential constructions, their intermodal lags were systematically briefer than those of habitual sequential integrators. As a result, an individual senior's integration pattern could be identified quickly and reliably, with temporal thresholds for fusion of their multimodal constructions adjusted accordingly. In this case, the temporal thresholds for simultaneous integrators could be adapted to be substantially shorter than those for sequential integrators, which in turn could increase the response speed, interactive synchrony, robustness and overall usability of the multimodal interface. Future work needs to evaluate alternative strategies for adapting temporal thresholds in time-sensitive multimodal architectures, as well as documenting specific performance advantages and possible tradeoffs.

In terms of human performance, seniors' task-critical errors, response latencies to execute a task, and self-regulatory language all increased progressively as the spatial complexity of tasks increased from low to high. In fact, 87% of seniors made task-critical errors at some point during their task, with the average percentage of such errors increasing from 4.4% to 10.3% to 25.7% between low, moderate, and high difficulty tasks, respectively. From a methodological viewpoint, this increasing rate of errors validates the calibrated increases in spatial difficulty levels. In addition, Figure 5 shows that increases in task difficulty influenced seniors' intermodal lags, which increased from .99 to 1.46 secs between the low and moderate/high difficulty tasks, or by 47%.

Given the combined effects of decreased motor and processing speed, diminished memory capacity, and challenging spatially-oriented tasks, it was expected that seniors would require considerable time to plan and initiate their tasks, and also would produce elevated rates of self-regulatory speech as they attempted to process and retain spatial information during map tasks. In fact, over 80% of the seniors engaged in preparatory self talk at some time during their session, which is considerably higher than the 10% observed in previous research for younger adults (S. Oviatt, personal communication, 2003). The likelihood that seniors would engage in self talk also increased steadily from 26.9% to 38.5% to 43.7% between low to high difficulty tasks, as shown in Figure 6. Furthermore, this increase in self talk was strongly correlated with the presence of human performance errors. Figure 7 illustrates that self talk began increasing when performance errors exceeded 10%, with a steep rise above 18% errors. Overall, a substantial 44% of the variance in self talk could be accounted for by knowing an individual senior's level of task-critical errors.

The following is an illustration of one task interaction taken from a senior's transcript:

> **System prompt:** Draw a sandbag fortification along the east waterfront from Broadway Bridge to Fremont Bridge.
>
> **User's self talk:** "Draw a sandbag fortification along the east water front from Broadway Bridge to Fremont."
>
> *(while looking at instruction)*
>
> **User's self talk:** "Fremont… Broadway Bridge to Fremont… Draw a sandbag fortifa-, east water front."
>
> *(while looking at map)*
>
> **User's input to system:** "I'm drawing a sandbag fortification from, to an east Fremont Bridge from the Broadway Bridge."
>
> *(while incorrectly drawing line along the <u>west</u> waterfront from Broadway to Fremont Bridge)*

After receiving instruction, this user engaged in extensive self talk while reading and comprehending their instructions, and then again while finding locations on the map display. This example shows that, while planning the delivery of their system input, self talk focused heavily on spatial location and directional terms. In spite of this user's extensive preparatory self talk, their final input to the system was disfluent. In this case, it also contained an *east-west* spatial directional error, which was the most common type of human performance error observed in these tasks.

This conspicuously high rate of self talk in seniors poses a challenge for the design of open-microphone engagement for future spoken language and multimodal interfaces. In fact, even newer audio-visual approaches to microphone engagement could not distinguish self talk from intentional input to a system in cases where the user is looking at the system while speaking [8]. To function robustly for user groups like children and seniors, or for relatively difficult tasks, audio-visual microphone engagement would need to incorporate additional sources of information such as reduced speech amplitude or lexical repetition. Although a click-to-speak microphone engagement implementation circumvents these problems, it will not necessarily remain the preferred option for many mobile or pervasive multimodal interfaces in the future.

In conclusion, the results of this research have implications for the design of adaptive processing strategies appropriate for seniors' applications, especially for the development of temporal thresholds used during fusion in time-sensitive multimodal architectures. They also have implications for anticipating and adapting multimodal interfaces to support limitations in human performance, especially in user groups like seniors. The long-term goal of this research is the development of flexible and robust multimodal interfaces, which will be essential for supporting *full spectrum* multimodal interfaces that can accommodate major individual differences among users.

5. ACKNOWLEDGMENTS

We would like to thank Courtney Stevens and Pam Schallau for subject recruitment and early piloting, Kristy Hollingshead for scoring, Stephanie Tomko for second scoring, and our senior volunteers for participating in the study. Thanks also to Jim Ann Carter and Lesley Carmichael for graphics and editing assistance, and to members of CHCC for many insightful discussions. Table 1 and figures 1 and 2 reprinted with permission from ICSLP'02. This research was supported by DARPA and NSF Grant No. IIS-0117868.

6. REFERENCES

[1] Berk, L.E. Why children talk to themselves. *Scientific American*, 1994, 271(5), 78-83.

[2] Cohen, P.R., M. Johnston, D. McGee, S.L. Oviatt, J. Pittman, I. Smith, L. Chen & J. Clow. QuickSet: Multimodal interaction for distributed applications. In *Proc. of Multimedia'97*, 31-40.

[3] Comblain, A. Working memory in Down's Syndrome: Training the rehearsal strategy. *Down's Syndrome: Research and Practice*, 1994, 2(3), 123-126.

[4] Czaja, S.J. & C.C. Lee. Designing computer systems for older adults. In *Handbook of Human-Computer Interaction* (J. Jacko & A. Sears, eds.). LEA, NY, 2002, 413-427.

[5] Heckmann, M., F. Berthommier & K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP JASP*, 2002, 11, 1260-1273.

[6] Huhns, M. & G. Weiss. Guest Editorial. *Machine Learning* (special issue on Multiagent Learning), 1998, 33(2-3), 123-128.

[7] Illina, I. Tree-structured maximum a posteriori adaptation for a segment-based speech recognition system. In *Proc. of ICSLP'02*, 1405-1408.

[8] Iyengar, G. & C. Neti. A vision-based microphone switch for speech intent detection. In *Proc. of RATFG-RTS'01*, 101-105.

[9] Jacko, J.A., I.U. Scott, F. Sainfort, K.P. Moloney, T. Kongnakorn, B.S. Zorich & V.K. Emery. Effects of multimodal feedback on the performance of older adults with normal and impaired vision. *Lecture Notes in Computer Science (LNCS)*, 2003, 2615, 3-22.

[10] Kart, C.S., E.K. Metress & S.P. Metress. *Aging, Health and Society*. Jones and Bartlett, Boston MA, 1988.

[11] Kemper, S. & T.L. Mitzner. Language, production and comprehension. In *Handbook of the Psychology of Aging 5th Ed* (J.E. Birren & K.W. Schaie, eds.). Academic Press, San Diego CA, 2001, 378-398.

[12] Luria, A.R. *The Role of Speech in the Regulation of Normal and Abnormal Behavior*. Liveright, NY, 1961.

[13] Oviatt, S.L. Ten myths of multimodal interaction. *Communications of the ACM*, 1999, 42(11), 74-81.

[14] Oviatt, S.L., P.R. Cohen, M.W. Fong & M.P. Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *Proc. of ICSLP'92*, 2, 1351-1354.

[15] Oviatt, S.L., P.R. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson & D. Ferro. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 2000, 15(4), 263-322.

[16] 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.

[17] Salthouse, T.A. *Theoretical Perspectives on Cognitive Aging*. LEA, Hillsdale NJ, 1991.

[18] Swanson, H. L. What develops in working memory? A life span perspective. *Dev. Psychology*, 1999, 35(4), 986-1000.

[19] Wingfield, A. & E.A.L. Stine-Morrow. Language and speech. In *Handbook of Cognitive Aging 2nd Ed* (F.I.M. Craik & T.A. Salthouse, eds.). LEA, Mahwah NJ, 2000. 359-416.

[20] Xiao, B., C. Girand & S.L. Oviatt. Multimodal integration patterns in children. In *Proc. of ICSLP'02*, 629-632.