Multimodal Integration Patterns in Children

Benfang Xiao, Cynthia Girand, and Sharon Oviatt

Department of Computer Science and Engineering
Oregon Health & Science University
+1-503-748-1342; {benfangx, girand, oviatt}@cse.ogi.edu
http://www.cse.ogi.edu/CHCC

ABSTRACT
Multimodal interfaces are designed with a focus on flexibility, although very few multimodal systems currently are capable of adapting to major sources of user or environmental variation. The development of adaptive multimodal processing techniques will require empirical guidance on modeling key aspects of individual differences. In the present study, we collected data from 24 7-year-old children as they interacted using speech and pen input with an educational software prototype. A comprehensive analysis of children’s multimodal integration patterns revealed that they were classifiable as either simultaneous or sequential integrators, although they more often integrated signals simultaneously than adults. During their sequential constructions, intermodal lags also ranged faster than those of adult users. The high degree of consistency and early predictability of children’s integration patterns were similar to previously reported adult data. These results have implications for the development of temporal thresholds and adaptive multimodal processing strategies for children’s applications. The long-term goal of this research is life-span modeling of users’ integration and synchronization patterns, which will be needed to design future high-performance adaptive multimodal systems.

1. INTRODUCTION
One pervasive theme of contemporary computing is the development of increasingly flexible and reconﬁgurable systems, examples of which include the design of multi-agent systems [1], mobile robotics [7], and multimodal interfaces [5]. This research is motivated largely by the desire to build more reliable or fault-tolerant systems for real-world usage contexts. In the case of multimodal interfaces, an additional goal is to design systems that are capable of accommodating a wide range of users. To achieve this goal, advanced multimodal interfaces will need to adapt to individual differences and variation in user state in order to become full spectrum interfaces.

To date, very few existing multimodal systems include any adaptive processing. One concept of adaptivity involves adaptation of the temporal thresholds used in time-sensitive multimodal architectures for determining when two input signals should be interpreted jointly. State-of-the-art systems like QuickSet use temporal thresholds based on empirically-derived user synchronization patterns and intermodal lags [5], but these temporal constraints currently are fixed rather than adaptive. The development of intelligently adaptive processing strategies will require data collection in key areas such as lifespan modeling of individual differences.

Previous research on the pen/voice multimodal integration patterns of adult users has reported that they are classifiable into two basic subtypes. Adults either consistently deliver pen input followed by speech after a brief lag as a sequential pattern, or else they deliver signal input in a temporally-overlapped or simultaneous manner. The majority of adults (i.e., 64%) tend to be sequential integrators, with pen input completed before speech begins [3, 6]. In these cases, the typical intermodal lag between the end of pen input and the onset of speech ranges between 0.0 and 4.1 seconds, and averages 1.4 seconds [6]. Whatever their predominant integration pattern, previous research has revealed that they will use it consistently for 90% of all their multimodal constructions [3]. Furthermore, their preferred pattern will almost always be predictable on their very first multimodal utterance [3].

In many respects, these data on individual differences in adult multimodal integration patterns present an ideal set of circumstances and opportunity for adaptive processing. That is, users are divided into two basic types, with early predictability and a high degree of consistency in their integration pattern. However, to date, empirical information on the precedence relations and intermodal lags between pen and speech input has been used mainly: (1) to determine the likelihood that an individual signal should be interpreted in a stand-alone manner, rather than waiting for an additional signal piece to interpret their combined meaning jointly, and (2) to establish fixed temporal thresholds for determining how long to wait for an additional signal piece, given that joint semantic processing appears likely and necessary.

However, future research still needs to model user groups besides adults, and to formulate and test adaptive processing strategies that could substantially improve multimodal system performance.

2. GOALS OF THE STUDY
The focus of the present research was a comprehensive analysis of children’s speech and pen-based multimodal integration patterns, and a comparison of these integration and synchronization patterns with those of adults. Such data are needed to establish accurate temporal thresholds for next-generation multimodal architectures that support children’s software applications, and also to begin designing effective adaptive strategies for a new class of advanced multimodal systems.

We hypothesized that children would:
• Spontaneously engage in multimodal interaction, with the likelihood increasing during error repair,
• Use speech and pen input to convey complementary semantic information.

With respect to synchronization of input modes, it was predicted that children would:
• Deliver pen input before speech,
Be classifiable as either simultaneous or sequential integrators,
Integrate modes simultaneously more often than adults,
Be predictable early during an interaction with respect
to their dominant integration pattern,
Remain consistent in their preferred integration pattern
throughout an interaction, and
Exhibit faster intermodal lags than adults during
sequential multimodal constructions.

3. METHOD

3.1 Participants, Task and Procedure

Twenty-four elementary-school children participated as paid
volunteers. The participants were evenly divided into two age
groups, younger children (mean age 8 years, 2 months), and older
ones (mean age 9 years, 7 months), with each group gender
balanced. Participation was conducted at a local school field site.

Children participating in the study were introduced to
Immersive Science Education for Elementary kids (I SEE!), an
application designed to teach them about marine animals and
graphing. The interface permitted children to use speech, pen or
multimodal input while conversing with animated software
characters as they learned about marine biology. The data
collection process was based on a simulation technique. Details of
the I SEE! simulation architecture, performance, and use in
research with children are described elsewhere [4]. Figure 1
illustrates the I SEE! interface.

During the study, children used the I SEE! interface to view
and interact with 24 different marine animals (e.g., octopus). The
marine animals were animated and available as “conversational
partners” who answered questions about themselves using TTS
output. An animated “Spin the Dolphin” character, shown in the
lower right corner of Figure 1, also was available as a
conversational partner. Spin answered the children’s questions,
assisted as needed with the interactive activity (e.g., controlling
videos), and provided entertainment (e.g., telling jokes).

Before starting a session, each child received instructions and
practice with a professional science teacher on how to use the I
SEE! interface on a small hand-held computer. The children were
told that they could use pen input in the white areas of the tablet,
tap and speak to the computer, or speak and use pen input
simultaneously. They were encouraged to use these input modes in
any way they wished, and multimodal interaction was never
specifically modeled for them. After the teacher left, the child
spent approximately one hour alone in a classroom using the
educational software. During this time, the child conversed with
24 marine animals, collecting information and building graphs that
represented information about the sea creatures (e.g., “Is this
animal poisonous?”). Children were encouraged to have fun
learning about the animals.

3.2 Transcription, Data Coding and Inter-coder Reliability

All interaction was videotaped. Children’s verbatim spoken
and written input was transcribed, as well as system responses. In
total, 17% of the intermodal lags were second-scored by an
independent coder. The average departure between coders was
0.05 seconds, and 80% of the lags matched to within 0.1 seconds.

<table>
<thead>
<tr>
<th>Pen Input</th>
<th>Speech Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. &quot;I don’t know&quot; (Child scribes, then responds verbally to Spin.)</td>
<td></td>
</tr>
<tr>
<td>2. “Food” (Child writes word, then speaks to Creature.)</td>
<td></td>
</tr>
<tr>
<td>3. “What’s one plus one equal?” (Child asks question, then writes equation for Creature.)</td>
<td></td>
</tr>
<tr>
<td>4. “Yes, it is true.” (Child writes “ye,” deletes it, then responds verbally.)</td>
<td></td>
</tr>
<tr>
<td>5. “Do you have a girlfriend?” (Child asks Creature question, then draws heart.)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Examples of qualitatively different multimodal utterances from the corpus with transcriber comments

4. RESULTS

In total, 8196 child utterances were transcribed during
conversational interaction with the computer, of which 6436 were
suitable for analysis in this research. Figure 2 illustrates five
eamples of multimodal utterances involving different pen content.

4.1 Multimodal Preferences and Individual Differences

Of the utterances analyzed, 651 (10.1%) involved multimodal
speech and pen input, 5165 (80.3%) were speech-only utterances,
and 620 (9.6%) were pen-only utterances. Each of the 24 children
constructed at least one multimodal utterance during their session.
They averaged 27.1 multimodal utterances per session, with
individuals ranging between 1 and 147. Figure 3 illustrates
individual differences in the percentage of utterances that children delivered multimodally.

Of the 651 multimodal utterances, 609 (93.4%) contained abstract scribbling with no apparent semantic meaning, 26 (4.0%) contained words, 17 (2.6%) contained symbols (e.g. checkmarks, digits), 3 (0.5%) contained gestures (e.g. deletion mark), and 2 (0.3%) contained pictures (e.g., drawing of fish).

4.2 Complementarity of Speech and Pen input

Of the 44 multimodal utterances with pen input that could be interpreted semantically, 41 (93.2%) involved speech and pen input with complementary semantic meaning (i.e. partially or completely non-overlapping semantics), and 3 (6.8%) involved completely duplicated speech and pen input.

4.3 Multimodal Input during Error Handling

A disfluent multimodal utterance is illustrated in Figure 2, Example 4. Of 80 spontaneous child disfluencies observed during multimodal constructions, a large percentage (22.5%) involved cross-modal repairs in which the user switched to a second input mode. Analyses revealed that children were significantly more likely to interact multimodally when they were disfluent and needed to repair an utterance (i.e., 13.9% multimodal constructions in disfluent utterances) than when no disfluency occurred (9.9% multimodal constructions), Wilcoxon Signed Rank test, \( T^+ = 25, N = 7, p < .04, \) one-tailed.

4.4 Simultaneous and Sequential Integration Patterns

Children constructed simultaneous multimodal utterances more often than sequential ones. Of the total, 451 (69.3%) utterances were simultaneous, and 200 (30.7%) were sequential. In addition, all children were classified as being either simultaneous or sequential integrators when 65% or more of their multimodal constructions predominantly fit one pattern. As shown on the left side of Figure 4, of the 13 children with at least 10 or more multimodal constructions, 10 (76.9%) were predominantly simultaneous integrators, and 3 (23.1%) were sequential integrators. All 13 children had a dominant integration pattern, and 12 of the 13 children adopted this pattern on their very first multimodal construction. Furthermore, their average consistency in maintaining that pattern during an interaction was 93.5%.

The right side of Figure 4 compares the child data with previous adult data [3, p. 78]. Children were significantly more likely to be simultaneous integrators than adults, as revealed by a Wilcoxon-Mann-Whitney test, \( z = 2.03, p < .025, \) one-tailed.

4.5 Temporal Precedence of Input Modes

Of 200 sequential multimodal utterances, 194 (97%) involved pen input preceding speech, with only 6 (3%) involving speech preceding pen input.

4.6 Intermodal Lags during Sequential Integrations

Figure 5 illustrates the distribution of intermodal lags during all of children’s sequentially-integrated multimodal constructions, which averaged 0.75 seconds. Of these, 60% ranged between 0.0 and 0.7 seconds, 87% between 0.0 and 1.4 seconds, and 100% between 0.0 and 2.1 seconds. Children’s lags during disfluent constructions were significantly longer than those during fluent constructions (mean 1.02 vs. 0.71 seconds, respectively), independent t-test \( t = 2.6 (df = 171), p = .005, \) one-tailed.

In contrast, Figure 6 reveals the intermodal lags for children’s sequential multimodal constructions involving only meaningful pen content (i.e., words and symbols, but not abstract scribbles). These child lags averaged 1.1 seconds, and ranged between 0.0 and 2.1 seconds. This average is comparable to average adult lags of 1.1 seconds for matched multimodal constructions taken from a previous study [9, p. 420], but adult intermodal lags ranged more widely between 0.0 and 4.1 seconds. These child lags during meaningful content were significantly longer than those for non-meaningful pen content (i.e., abstract scribble only), with means of 1.05 vs. 0.71 seconds, respectively), independent t-test, \( t = 3.05 (df=171), p < .0015, \) one-tailed.
5. DISCUSSION

The pen/voice multimodal integration patterns reported for children and adults are strikingly similar in many ways. Individual users in both groups can be classified as either simultaneous or sequential integrators, and their predominant pattern is predictable early and remains highly consistent. In the case of children, the average consistency of individual integration patterns during a human-computer interaction was 93%, and for 12 of 13 children their dominant pattern was predictable on the very first multimodal construction. This high degree of consistency may have reflected the adoption of a "success strategy," although further research is needed to clarify the underlying cause. In comparing sequential multimodal constructions matched on meaningful semantic content, the average intermodal lag was 1.1 seconds for both children and adults. Finally, both children and adults combined speech and pen input that conveyed complementary semantic information, and there is a strong tendency for pen input to precede speech. In fact, during children’s sequentially-integrated multimodal utterances, pen input preceded speech 97% of the time. All of these results generalized across age groups, in spite of the fact that the specific applications (i.e., service transaction vs. educational exchange), dialogue models (i.e., command vs. conversational), and interface feedback (i.e., graphical vs. multimedia) varied in these different studies.

In contrast, children were significantly more likely than adults to integrate their multimodal constructions simultaneously. In fact, 77% of children were classifiable as simultaneous integrators, compared with only 36% of adults. In addition, children’s intermodal lags ranged consistently faster than adults’ for all of the data analyzed: 0-2.1 seconds, compared with 0-4.1 seconds. Finally, children clearly were more manually active than adults, which was evident in their high rates of abstract scribbling. An analysis of all of children’s multimodal constructions, including those in which pen input involved scribbling and graphics, revealed that children’s intermodal lags averaged faster than those of adults— 0.75 versus 1.1 seconds, respectively. Further analyses revealed that children’s lags were faster when pen input was not meaningful (e.g., scribble), and also when their multimodal constructions were not disfluent.

Children’s simultaneous integration patterns and consistently faster lags may in part reflect faster motor speed and greater behavioral impulsivity. The results reported in this research indicate that temporal thresholds established for multimodal architectures supporting children’s software applications could be reduced to a maximum 2-second delay, which is half the wait required for adult users. In addition, since more children are consistent simultaneous integrators than adults, future adaptive interfaces could be designed to frequently eliminate wait times altogether for this population.

In conclusion, empirical modeling of the type described in this paper is expected to yield new strategies for adapting the temporal thresholds of time-sensitive architectures, which in turn can lead to substantial improvements in multimodal system response speed, the interactive synchrony of the interface, and overall system robustness. The long-term direction of this research is the development of flexible and robust multimodal interfaces, which will be essential for supporting full spectrum interfaces that accommodate major individual differences among users.

6. ACKNOWLEDGEMENTS

This research was supported in part by Grants No. IRI-9530666 and IIS-0117868 from the National Science Foundation, Special Extension for Creativity (SEC) Grant No. IIS-9530666 from NSF, and a gift from the Intel Research Council. We are grateful to Matt Wesson for programming and testing, Rachel Coulston and Courtney Darves for transcription, and our student volunteers.

7. REFERENCES