Designing and Evaluating Conversational Interfaces with Animated Characters

Sharon Oviatt and Bridget Adams

X.1 Introduction: Ingredients of a Conversational Interface

During the past decade, due largely to progress inspired by the DARPA Speech Grand Challenge project and similar international efforts (Martin et al. 1997; Cole et al. 1997), significant progress has occurred in the development of spoken language technology (SLT). Spoken language systems now are implemented extensively for telephony applications (Spiegel and Kamm 1997), as well as on workstations and even small palm computers. They also are supporting new training systems for learning foreign languages and basic reading skills, as well as the commercialization of numerous automated dictation systems for applications such as medical charting, legal records, and word processing. The rapid maturation of spoken language technology, along with advances in natural language processing, also has stimulated interest in the development of increasingly conversational interfaces.

The emerging interest in conversational interfaces presents a particularly intriguing but challenging research agenda. In contrast to alternative styles of spoken language interface, including command language and dictation, conversational interfaces aim to support large-vocabulary spontaneous spoken language to be exchanged as part of a fluid dialogue between user and computer. They also permit the user to take extensive initiative in shaping this dialogue. To accomplish these goals, conversational interfaces support parallel natural language processing for both input and output between human and computer.

However, several other elements of human “conversation” exist that make conversational interface design a challenging direction. First, human conversation is not only an information exchange but also an informal social activity that is inherently reinforcing. In fact, a “conversationalist” is a person who enjoys and contributes to good conversation. As a result, conversational interfaces are considered “social interfaces,” and when we participate in them we respond to the computer at some level as a social “persona” or partner (Nass, Steuer, and Tauger 1994). This recognition, in part, has led to the design of new animated software characters, which basically serve as an interface vehicle to facilitate human-computer conversational exchange. The new field of research on
animated characters has become an unusually diverse and generative area, as is evident from the range of content in this book.

Another challenging element of human “conversation” that currently is driving conversational interface design is the interest in supporting more transparent, high-bandwidth, and relatively natural multimodal communication (Oviatt et al. n.d.). This movement aims to process human input from two or more modes like speech, pen, touch, gesture, and gaze, while simultaneously generating coordinated multimedia output from text, graphics, animation, speech and nonspeech audio—or their combination in the form of lifelike animated characters (André and Rist this volume; Badler et al. this volume; Ball and Breese this volume; Cassell et al. this volume). Such multimodal communication is viewed as a more powerful and expressive means of human-computer interaction, in addition to being well suited to the design of richer and more credible conversational interfaces.

The design of effective conversational interfaces clearly is a wide open research topic. It also is one that relies heavily on multidisciplinary expertise and teamwork, which constitutes a challenge in itself. To design conversational interfaces that function effectively, we need to start at the beginning—by asking a series of deceptively simple questions. Why do we want conversational interfaces in the future? How can we design them to produce the landmarks of a stimulating conversation—engagement, responsivity, reciprocity, synchronization, focus, intelligibility? And how will we evaluate that we have achieved these remarkable but elusive goals?

In this chapter, we introduce a new Mobile Simulation Environment, which was developed to support advance data collection and the design of conversational interfaces involving (1) new media (e.g., speech) and multimodal input, (2) portable technology for use in field settings and in situations where users are mobile, (3) diverse user groups (e.g., children), and (4) different animated character concepts. We also describe the results of an initial study based on data collection with this simulation tool and the I SEE! (Immersive Science Education for Elementary kids) educational application. In this study, the spoken language of ten 6- to 10-year-old children was compared while interacting with an animated character, and again while interacting with a human adult partner.

The specific goal of this research was to investigate hard-to-process articulatory, lexical, and grammatical features of children’s speech, such as spoken disfluencies, which may require specialized language modeling and interface design to build successful interactive systems. Fragmented and disfluent
language are known to be particularly challenging features for any spoken language system to process, even in adult language, so the evaluation of children’s disfluent speech was a major aim of this research. Since educational technology typically is used in field settings (e.g., school, home) and may best be applied when children are immersed in natural learning opportunities (e.g., marine biology center), the Mobile Simulation Environment was developed as a tool for collecting speech samples in realistic usage contexts. Due to noise levels, collaborative interaction and other factors, the rate of fragmented and disfluent speech would be expected to be higher in such field contexts than in an artificial laboratory one. The general long-term goal of this research is the development of robust conversational interfaces for children, especially ones that can be used in natural environments, and that make effective use of animated characters for purposes like education.

X.2 Tools for Designing and Evaluating Conversational Interfaces

X.2.1 Simulation Techniques

Simulation studies, also known as wizard-of-Oz studies, have become an important and flexible tool for investigating human-computer interaction, and for designing systems that are still in the planning stages. These studies involve proactive and situated data collection for the purpose of system design, which is done in advance of actually building a fully functional system. In a semiautomatic simulation, a test subject uses what she believes is a fully functional system, while a programmer assistant at a remote location provides simulated system responses. As the subject works, the assistant tracks the subject’s input and simply clicks on predefined fields on his or her workstation to send system confirmations back to the subject. Recently, high-fidelity simulation environments have offered an ideal means of conducting advance empirical work on spoken language and multimodal systems.

A high-quality simulation typically includes extensive automation in order to support accurate and rapid responding by the programmer assistant. For example, simulated recognition errors can be delivered automatically by a random error generator, which is preprogrammed to deliver different base rates and types of recognition error distributed randomly across task content (Oviatt 1996). Simulations also aim to support speedy interaction, especially for interfaces that include spoken language. In this type of simulated interface, an emphasis typically is placed on supporting subject-paced interactions that average less than a one-second delay between the subject’s input and system response (Oviatt et al. 1992).
One reason that high-fidelity simulations are the preferred method of designing and evaluating planned systems is because they are relatively easy and inexpensive to adapt, compared with actually building and iterating a complete system. They permit researchers and system designers to alter a planned system’s characteristics in major ways (e.g., the input and output modes available in a multimedia system) and to study the impact of important interface characteristics in a systematic and scientific manner (e.g., the type and base rate of system errors). In comparison, any particular system with its fixed characteristics is a less flexible and suitable research tool. In a practical sense, simulation research can assist in the evaluation of critical performance tradeoffs and in making decisions about alternative system designs, which designers must do as they strive to create more usable spoken language systems. Using simulation techniques, rapid adaptation and investigation of planned system features permit researchers and system designers to gain a broader and more principled perspective on the potential of whole newly emerging classes of technology, including next-generation conversational interfaces, multimodal/media systems, and animated character design.

Using simulation techniques, researchers are able to collect empirical data that can

- reveal undiscovered phenomena of interest, such as landmark features of interactive speech, that will need to be processed by future spoken language and multimodal systems
- quantify the prevalence of linguistic and other behavioral phenomena observed in users
- establish their causal basis through isolation and manipulation of the factors that drive them
- interpret these linguistic and behavioral phenomena in relation to contextual factors that predict and explain them
- create a solid foundation for next-generation interface design

X.2.1 Mobile Simulation Environment

One limitation of previous simulation environments is that they have focused on supporting laboratory-based research rather than real-world usage contexts. In addition, they have not been tailored to explore interface designs appropriate for a diverse range of more “challenging” users, such as children. Finally, they have yet to be applied to the investigation and design of next-generation conversational interfaces that incorporate animated characters.
In recent years, interest has grown in developing small handheld systems with more natural and powerful interfaces that can be used in varied real-world usage contexts by diverse users. Partly as a result of this trend, the goal of the new Mobile Simulation Environment, which was developed for this research, is to create a high-fidelity tool to support advance data collection and interface design appropriate for (1) new media (e.g., speech, pen, touch) and their multimodal combination, (2) portable technology for use in natural field environments and in situations where users are mobile, (3) diverse user groups (e.g., children or elderly users, accented nonnative speakers), who represent significant market potential, and (4) conversational interfaces that incorporate animated characters and other multimedia output. To further promote these types of future system design, an additional goal of the Mobile Simulation Environment is to develop (5) a toolkit that enables the rapid creation and adaptation of different simulated applications.

One example of an application currently being developed in conjunction with the Mobile Simulation Environment is Immersive Science Education for Elementary kids-I SEE! This application teaches early elementary children about marine biology and simple data tabulation and graphing concepts. Its features include (1) rich National Geographic video segments that kids can query for information, (2) natural conversational interaction with animated video content as well as designed animated characters, (3) control over the video display, and (4) data tabulation and creation of graphic displays. This functionality also is designed to be (5) controlled by children's natural multimodal spoken and pen-based input, and (6) to be used on small handheld systems while children are immersed in natural learning opportunities—whether at school, at home, or passing by exhibits at a marine science center.

Figure X.1 illustrates the Mobile Simulation Environment’s testing setup. During data collection with I SEE!, a child interacts multimodally with the application using a handheld device—for example, while learning about the sea otters that he or she is observing outdoors at the marine science center (fig. X.1, top panel). During this simulation testing, the child believes that he or she is interacting with a fully functional system, although a programmer assistant actually provides system responses from a remote location as this user data is collected over a wireless network (fig. X.1, bottom panel). To ensure high-fidelity data capture from mobile users, "local calibration" of user's input (e.g., inking and speech characteristics) also is monitored directly from their portable device using a small wearable video camera.

Fig. X.1 here
This simulation data collection method is based on a distributed agent-building environment that uses a modified version of the Open Agent Architecture (OAA) (Cohen et al. 1994), COM, and DCOM. The component agents include (1) Microsoft “agent” characters, (2) video control, and (3) an inking agent. The basic architecture for the Mobile Simulation Environment is illustrated in fig. X.2, with the user’s display screen on the left and the linked wizard’s screen and control palette on the right. The subject’s user interface accommodates speech and pen-based input, which can be combined multimodally. The system’s multimedia output includes speech and nonspeech audio, text, graphics, video, and animated movements. The programmer assistant or wizard’s interface is organized with database-driven scenario control, and uses the main OAA agent architecture, COM, and DCOM components.

fig. X.2 here

The Mobile Simulation Environment runs on a PC using WIN 95 or NT. Using wireless networks, it can receive data from portable computing devices such as the Fujitsu Stylistic 1200 and DRS machines. Data collection with smaller CE devices also is planned for the near future.

X.3 Talking to Thimble Jellies: How Kids and Animated Characters Converse

X.3.1 Related Literature on Animated Characters

At present, very little is known about how real users will interact with animated characters, especially child users. Most studies performed to date have focused on adult users’ subjective evaluation of animated characters, especially ratings of their social and personality characteristics, general attractiveness, trustworthiness, persuasive appeal, and so forth (Nass, Isbister and Lee this volume; Lester et al. 1997; van Mulken, André, and Muller 1998). In a few studies, behavioral measures also have been evaluated to compare adult performance in an interface with and without an animated character. For example, in an interesting study, van Mulken, André, and Muller (1998) evaluated the impact of presenting an animated agent with text-to-speech output compared to presenting the application content via text and object animation. Although their subjects preferred the animated character and believed that it helped them concentrate better, it had no actual effect on people’s comprehension and recall of technical or nontechnical material. In other studies on educational impact, anecdotal and preference measures again have demonstrated that animated
characters are appealing and apparently motivating to students (Lester et al. 1997), although no firm behavioral evidence exists yet that they promote significant learning.

Among the few behaviorally oriented evaluations of animated characters, some have begun examining users’ communication patterns during human-computer interaction. For example, during development of the Baldi talking head, Massaro et al. (this volume) compared users’ perception of different speech sounds while listening to text-to-speech augmented with a view of Baldi’s lip movements to their perception of a natural voice. This work has entailed controlled evaluation of the impact of different versions of Baldi’s animation on the relative intelligibility of speech sounds, including a thorough comparison of the speech confusion matrices while listening to Baldi versus listening to a human speaker. The results of these numerous iterative evaluations have been used to refine Baldi. More generally, they also have advanced the state of the art on intelligible lip synchronization and animated conversational interface design.

In another recent study, Cassell and Thorisson (1999) investigated users’ general communicative responses while interacting with three different types of animated characters, including ones that (1) only supplied propositional feedback, (2) responded with propositional feedback and information on emotional state, and (3) responded with propositional and “envelope” feedback (i.e., nonverbal gaze and gestures). In this research, they examined users’ overlapped speech and total dialogue contributions. Their results indicated that the character with envelope feedback promoted the most fluid and efficient interaction. The character based on this design also was rated by users as higher on language ability.

Other recent studies have begun exploring users’ speech production while interacting with an animated character. For example, the Swedish August project has collected thousands of samples of human speech from public kiosk users during interaction with an animated character (Bell and Gustafson 1999a). Recent analyses have confirmed hyperarticulate speech adaptations in Swedish users’ spoken language to an animated character during error handling (Bell and Gustafson 1999b). This particularly extensive and diverse corpus currently is being analyzed in linguistic detail, so in the future we may learn considerably more about how people actually talk to animated characters.

Although a few recent studies have begun exploring users’ communication patterns while conversing with an animated character, nonetheless a great deal remains unknown about human communication with such interfaces.
Furthermore, the specific focus of previous work has yet to address children’s spoken language to animated characters. Previous language development literature indicates that children’s speech production is considerably more challenging than that of adults, because it is less mature, more variable, and dynamically changing from year to year (Lee, Potamianos, and Narayanan 1997; Shatz and Gelman 1973; Yeni-Komshian, Kavanaugh, and Ferguson 1980). Very recent research also has estimated that children’s speech is subject to recognition error rates that are two to five times higher than adult speech (Aist et al. 1998; Das, Nix and Picheny 1998; Potamianos, Narayanan, and Lee 1997; Wilpon and Jacobsen 1996). It therefore is clear that considerable corpus collection, language modeling, and tailored interface design will be needed before successful interactive technologies can be developed for children.

X.3.2 Goals of the Present Data Collection

This study compared the spoken language production of ten 6- to 10-year-old children interacting with an animated character in the I SEE! Interface with that of the same children interacting with a human adult partner. The investigation addressed the following general questions:

1. Will children engage in extended “conversations” with an animated character? If so, what kind of content do children initiate during these exchanges? And what are the characteristics of children’s speech and language with the animated character?

2. Even though we know adults speak differently to a “computer” than to another person, do children likewise speak differently to an animated computer agent? Or is their speech indistinguishable from the way they speak with a human partner? If their speech differs, then in what respect? Is it simplified or hyperarticulated, as an adult’s speech would be?

3. Is children’s speech to an animated character harder to process than adult speech to a computer? If so, what is more difficult, and how much more difficult is it?

4. Do children like interacting with animated characters? How lengthy are their interactions, and how absorbed do they become? Can animated characters be used as an effective interface tool for engaging children in conversational interfaces? What potential do they have for use in educational applications, and how should they be designed for such purposes?
Because so little past research has addressed children’s spoken language to computers, in this study we were particularly interested in analyzing the different and potentially hard-to-process articulatory, grammatical, and lexical features of children’s speech to an animated character. For example, disfluencies have been analyzed quite extensively in adult speech to computers, although to our knowledge they have yet to be evaluated in either adult speech to animated characters or in children’s speech at all. Disfluencies in spontaneous speech are self-repairs that interrupt the smooth flow of an otherwise coherent linguistic construction. They are widely recognized as presenting a major challenge and source of failure for current spoken language systems. Recently, researchers have become interested in modeling disfluencies as a landmark feature of spoken language, rather than an aberration or linguistic deficiency (Oviatt 1995; Shriberg 1994). Factors that predict disfluent speech also have been identified. Finally, corresponding interface designs have been formulated that are capable of substantially minimizing disfluencies (i.e., by up to 80%), although in a manner that remains completely transparent to users (Oviatt 1995).

For general reference, table X.1 summarizes estimates of the rate of adult disfluencies per 100 words in a variety of different types of spoken interaction. Among other things, it reveals that disfluencies occur at consistently higher rates in adult interpersonal speech than in human-computer speech. During adult human-computer communication, disfluencies also are known to bear a positive linear relation with utterance length, which accounts for approximately 80% of all the variance in spoken disfluencies (Oviatt 1995, 1997). This relation has been attributed in part to the increasing cognitive load associated with planning the content and order of information to be articulated, which rises progressively in longer utterances.

Table X.1 here

In this study, we were interested in comparing the rate of disfluencies in children’s speech with that previously estimated for different types of adult speech. For comparable types of spoken interaction, we hypothesized that child disfluency levels might be elevated. We also were interested in directly comparing the rate of disfluencies in children’s speech to an animated computer partner with that addressed to a human adult. If young children are mesmerized by animated characters and believe them to be “real,” then their disfluency levels should be indistinguishable. On the other hand, if children are engaged but still quite aware of speaking to a computer, then their data should resemble the adult pattern of higher interpersonal than human-computer disfluencies. A third goal was to evaluate whether utterance length bears the same strong linear relation to
children’s spoken disfluency rates, as it does in adults. Alternatively, if children experience greater cognitive load associated with planning utterances of increasing length, then the slope of their linear function may be steeper than that for adults. In addition to these specific articulatory issues, a further aim of the present study was to explore other potentially hard-to-process features of children’s speech, including invented or misapplied lexical content, immature grammatical constructions, and so forth.

X.3.2 Methods

X.3.2.1 Participants, tasks, and procedure

Ten children between 6 years, 8 months old and 10 years, 9 months old participated in this research, including seven females and three males. All were volunteers from the local Portland, Oregon community.

The study consisted of two data collection phases. During the first part, children interacted with the simulated I SEE! system (see sec. X.3.2.2) in a room set up as an informal children’s bedroom and playroom. The room was designed to be inviting and relaxing, with a sofa, stuffed animals, children’s artwork, a toy chest, and so forth. During the session, children sat on the sofa with a Fujitsu 1200 handheld PC angled in front of them, so that they could use pen input and view the screen comfortably. They were introduced to the I SEE! system by the experimenter, who spoke to them using a dolphin puppet. Each child received a ten- to fifteen- minute hands-on orientation to the system and its coverage, including how to (1) use speech and pen input, (2) make corrections, (3) control the movies, (4) use the system to learn about different animals, and (5) make a graph summarizing their findings. During the orientation, children also interacted with three practice animals and graphed information about them.

After this orientation, the experimenter left the children alone to play with the I SEE! system while they interacted with eight unique undersea animals (e.g., manatee, squid, stonefish). In each case, the children were told that they could ask the animals anything they wished but to be sure to ask Spin the dolphin’s question each time they met a new animal (e.g., “Are you poisonous or not?”). Spin was available as an animated guide to help the child with any questions about the application’s functionality, general interface controls, or graph construction. The children also were asked to enter the answer to Spin’s question on their individual graphs, which they could print and keep when done. A video record was made of the interface during the entire interaction, along with a closeup view of the children’s face and spoken language as they played. Following this interaction,
the experimenter returned and engaged the child in an informal discussion about what they thought of the I SEE! system and its animated characters. During some sessions when several children visited the lab together, an informal brainstorming and “design meeting” also took place between the children, experimenters, and programmers. The general purpose of such meetings was to collect and consolidate children’s design ideas during a playful group exchange in which they participated as the “experts” and “advisors” (Druin 1999).

During the second phase of the study, the same children participated in a game of Twenty Questions with the adult experimenter as they attempted to discover the identity of different animals. Children’s speech was audiotaped throughout this interpersonal interaction. In both of these phases of the study, children basically were involved in a gamelike interaction in which the goal was to extract information about the features or identity of different animals. For both tasks, children needed to take initiative in asking questions as they retrieved information and solved a simple problem. However, in phase one their resource was an animated software agent, and in phase two it was a human adult.

**X.3.2.2 I SEE! application and interface**

Figure X.3 illustrates I SEE!’s interface. The left side of the screen displays movies of marine animals. The child user could control the start, stop, replay, and automatic location of different marine animals on the movie by entering pen input in the white field beneath the movie, or by tapping the pen in this field and speaking. When the movie stopped, the marine animal was embellished with animated eyes that blinked naturally. At this point, the animal became available as a “conversational partner” for answering questions about itself. For example, an animated manatee could identify itself, its diet, habitat, unique behavior, endangered species status, and so forth. Essentially, animated eyes that gazed at the child provided attentional cues, and these cues were used to mark the transition from a passive movie-viewing experience to active availability of the animal as the child’s conversational partner.

Fig. X.3 here

The upper right side of figure X.3 illustrates a graph created by an eight-year-old child after querying six different animals about their endangered species status. After conversing with each animated creature and collecting information needed to classify it, the child entered each animal in her bar graph. In this case, she did so by entering pen input directly into the next available white field on the graph. The I SEE! interface permitted children to construct their graph in a direct
hands-on manner by using pen input to place symbolic artwork, animal names, or marks on the correct side of the graph or by pointing to a field and speaking their input. The system acknowledged the child’s entry by painting the field blue and making a splashing sound as the graph became taller on that side.

The lower right corner of figure X.3 also illustrates the animated dolphin character named “Spin.” The dolphin provided motivation and guidance by answering questions, making suggestions, issuing praise and reminders, and telling jokes. The dolphin was animated with natural eye movements, lip synching, facial expressions, and gestures. Spin also made exuberant large body movements, such as jumping, somersaulting, diving and splashing, and clapping. During conversation, Spin was capable of acknowledging her understanding, or expressing puzzlement and lack of comprehension, through her facial expressions and gestures.

During interaction, both the dolphin and marine animals responded using text-to-speech (TTS). The voices used for this application were female and male ones from Microsoft’s Whisper TTS version 4.0, with some hand-tailoring of pitch, duration, volume, intonation, and pronunciation. The voices were crafted to optimize intelligibility and character individuation, which are known to be influential in shaping a user’s reactions to an interface (Nass, Steuer and Tauber 1994). The TTS also was pilot tested for intelligibility with children and modified as needed.
X.3.2.3 Transcription and data coding

   Each participant’s speech was transcribed from the videotapes collected during their human-computer session and from the audiotapes during their interpersonal session. Transcription was done by a native speaker of English and was second scored for reliability. Attention was paid to transcribing verbatim spoken input, without “cleaning it up” in any way. This included recording spoken language phenomena such as nonword sounds, repetitions, disfluencies and self-repairs, confirmations, and so forth. Children’s speech in the transcripts was divided into two categories: (1) simple responses to queries elicited by the system (e.g., System: “Would you like to hear another joke?” Child: “Yes, okay.”), versus (2) child-initiated spoken dialogue (Child: “Do you change colors rapidly?” System: “No, I stay the same color.”)

   • Total words- The total number of child-initiated spoken words was tabulated for each child during both human-computer and interpersonal communication. Since these tasks were driven primarily by the child, less than 5 percent of all spoken language involved simple elicited responses to the system, and these data were excised for analytical purposes. Data on the total number of spoken words primarily provided a baseline for converting disfluencies to a rate per 100 words.

   • Mean length of utterance (MLU)- The average number of spoken words per utterance was tabulated for each subject during human-computer and interpersonal communication. These averages also were based on child-initiated spoken dialogue. MLU primarily was used for examining the relation between utterance length and disfluency rate in children’s speech.

   • Utterances per dialogue turn- The average number of utterances per dialogue turn, or before the system made a dialogue contribution, was tabulated for each child during human-computer and interpersonal communication. This measure in part reflects the degree of interactivity during the conversational exchange.

   • Requests for repetition- The number of times that each child asked his or her human or animated conversational partner to repeat an utterance was totaled, and a percentage of requested repetitions out of the total utterances for that child’s interaction then was calculated. This measure was used as an index of the relative intelligibility of the text-to-speech (TTS) output used in I SEE!

   • Disfluencies- Spontaneously occurring spoken disfluencies were totaled for each subject during both the human-computer and interpersonal communication sessions, and then were converted to a rate per 100 words. Disfluencies were
scored only during child-initiated spoken language, not during utterances involving system-elicited responses. In addition, only spontaneously occurring disfluencies were scored, not cases in which a repetition was elicited by a system interruption or an unusually slow system response. The following types of disfluency were coded: (1) content self-corrections, meaning errors in task content that were spontaneously corrected as the child spoke (e.g., “What kind of fish is that spotted, or no, striped one?”), (2) false starts, or alterations to the grammatical structure of a spoken utterance (e.g., “How many– do you have live babies or lay eggs?”), (3) repetitions—namely, repetitions of a phoneme, syllable, word, or phrase that occurred spontaneously as the subject spoke (e.g., “Are you endangered?”), and (4) filled pauses, those spontaneous nonlexical sounds that fill pauses in running speech (e.g. “Uh, how long does a sea turtle live?”), often signaling the start of a new phrase or self-correction. Scoring included cases in which multiple instances of different kinds of disfluencies occurred within a single utterance, as in “How– do you– do you know how many, uh– different species you have of your kind?” For further classification and coding details, see Oviatt 1995. For the purpose of this study, no attempt was made to code minor spoken mispronunciations like elongated or slurred sounds, interjection or omission of individual vowel and consonant sounds, and so forth, which are more difficult to identify reliably during scoring.

• **Idiosyncratic linguistic constructions**- Several types of linguistic irregularity were coded and totaled for each child during the human-computer and interpersonal communication sessions, and then were converted to a rate per 100 words. These linguistic idiosyncrasies were scored only during child-initiated spoken language, and they included the following phenomena: (1) invented words (e.g., “You’re the amazingest animal I’ve ever seen. Are you a tape-eater?”), (2) incorrect lexical selections (e.g., “Are you endangered or instinct?”), (3) concatenated words (e.g., “What are krill-shellfish?”), (4) ill-formed grammatical constructions (e.g., “Have you ever ate spaghetti?”; “You’re the most smartest fish.”), and (5) distinctly mispronounced or exaggerated articulations, sometimes exclamatory or highly emotive in nature (e.g., “What are mullooskies?”; “Are you a stuffed aminal?”; “Eeeeh!”; “Whooooaa!”).

• **Reliability**- A second scorer independently checked 100 percent of the transcriptions and disfluency codings from both studies, and any discrepancies were resolved before analyses were performed.

**X.3.3 Results**
The data yielded a child speech corpus totaling over 9600 words, including over 500 spontaneous disfluencies for analysis.

Figure X.4 illustrates that the same children had significantly higher levels of spontaneous speech disfluencies during interpersonal communication than while conversing with animated characters, paired \( t \)-test, \( t = 4.55 \) (\( df = 9 \)), \( p < .0005 \), one-tailed. The children’s disfluency rate averaged 6.71 percent during interpersonal communication, compared with 2.32 percent during human-computer communication, which represented almost a threefold higher rate during interpersonal communication. Figure X.4 also clarifies that all ten children had a higher rate of interpersonal disfluencies than human-computer disfluencies. Furthermore, the ranges were very distinct, with the rate of human-computer disfluencies ranging from 1 to 4 percent, and interpersonal disfluencies ranging from 4 to 12 percent. A median-split comparison of children’s disfluencies as a function of age (i.e., 6–8 versus 8–10 years) did not reveal any significant change in either human-computer or interpersonal rates within this age range.

Fig. X.4 here

Analyses revealed that the ratio of children’s speech comprised of verbatim utterance repetitions averaged 20 percent during human-computer communication, but only 7 percent during interpersonal interaction. That is, children’s speech directed to animated characters was significantly more repetitive than was their interpersonal speech, paired \( t = 3.62 \) (\( df = 9 \)), \( p < .003 \), one-tailed. Since repeated utterances may not require as much cognitive load during utterance planning, children’s disfluency rates were recalculated after removing all verbatim repetitions. However, this analysis reconfirmed that children’s disfluency rates were still significantly higher during interpersonal communication than while talking to an animated character, paired \( t = 3.99 \) (\( df = 9 \)), \( p < .0015 \), one-tailed.

Analyses of these data revealed that children’s utterance length averaged 4.8 during human-computer interaction and 4.3 during human-human interaction, which was not a significant difference by paired \( t \)-test, \( t = 1.47 \) (\( df = 9 \)), N.S. Although utterance length is known to correspond to disfluency rates, it clearly did not account for the differences between human-computer and interpersonal disfluency rates observed in these data. A comparison of the number of utterances that children spoke per dialogue turn also revealed no difference when speaking to an adult versus an animated character, paired \( t < 1 \). In these structural respects, then, children’s spoken dialogue with an animated character was similar to that with a human adult partner.
Further analyses revealed that children’s disfluency rate was higher than that previously reported for adults during unconstrained human-computer interaction involving a visual-spatial task, paired $t$-test, $t = 2.35$ ($df = 10$), $p < .025$, one-tailed. These comparisons involved child and adult utterances matched on MLU, since children’s MLU is shorter than that of adults and MLU is highly correlated with disfluency rates. After matching on utterance length, in this study the overall disfluency rate averaged 64 percent higher for children than the rate previously reported for adults.

A regression analysis indicated that the strength of predictive association between utterance length and spoken disfluency rate for 6- to 10-year-old children was $\rho^2_{XY} = .88$ ($N = 11$). That is, 88 percent of the variance in the rate of spoken disfluencies for children was predictable simply by knowing an utterance’s specific length. This is similar, but slightly higher than, the 79 percent strength of predictive association documented previously for adults in a map-based visual-spatial domain. The linear function illustrated in the scatter plot in figure X.5 summarizes this relation for both children and adults. For children, the $y$-axis constant coefficient is $-0.41$, and the $x$-axis beta coefficient of $+0.39$ represents utterance length. Children’s regression line is the steeper slope shown on the left in figure X.5. Their disfluency rate rises above 3 percent after an MLU of 7 has been reached, in comparison with an MLU of 12 for adults.

Fig. X.5 here

The level of idiosyncratic linguistic constructions in children’s speech averaged 1.83 percent overall for this age group, with no significant difference between human-computer (2.42%) and interpersonal speech (1.24%) in terms of rate of occurrence. A median-split comparison of the rate of idiosyncratic language as a function of age (i.e., 6–8 versus 8–10 years) revealed a significantly higher level in 6- to 8-year-olds than 8- to 10-year-olds during human-computer interaction, paired $t = 3.12$ ($df = 8$), $p < .01$, one-tailed. Figure X.6 illustrates the cumulative percentage of children’s spoken language that contained hard-to-process disfluencies or idiosyncratic language, which averaged about 5 percent of all human-computer communication and 8 percent of interpersonal communication during the age range studied.

Fig. X.6 here

Although children never requested that a human dialogue partner repeat an utterance, half of the children did request at least one repeat when interacting with
an animated character that used TTS output. However, only 1.5 percent of all utterances delivered by an animated character resulted in the child requesting a repetition. This indicates that the TTS was adequately intelligible, and children were infrequently distracted by it during their conversation.

X.3.4 Discussion

The following conversation between eight-year-old Edward and the thimble jellyfish demonstrates that children can indeed become very engaged while interacting with animated agents:

Edward: Stop! Hello. What are you? [videotape stops; jellyfish looks at Edward and blinks]
Thimble jelly: I am a thimble jellyfish.
Edward: Cool. Are you poisonous?
Thimble jelly: Yes, I am poisonous.[Edward draws a fish on the poisonous side of his bar graph]
Edward: You kinda look like a jellybean. What is your name?
Thimble jelly: I am Tiny the thimble jelly. I swim with hundreds of other thimble jellies.”
Edward: Are you common or exti– instinct?
Thimble jelly: I am very common.
Thimble jelly: I eat plankton. Sometimes sea turtles eat me.
Edward: Whoa, that’s not very much. All I eat is spaghetti, hamburgers. I won’t touch fish.
Thimble jelly: Yummy.
Edward: Have you ever ate spaghetti?
Thimble jelly: No way, José!
Edward: Are you a girl or a boy?

During these conversations, children spoke directly to the animals using personal pronouns, and approximately one-third of all the content exchanged involved social questions initiated by the child about the animated character’s name, birthday, personal characteristics, friends, and family life. Children typically reported that they enjoyed interacting with I SEE! They often specifically mentioned that they liked “talking to the animals,” whom they viewed as “friends” rather than parents or teachers. As expected, however, most of the remaining two-thirds of the conversational content focused on elementary marine biology information.
Edward’s conversation with the thimble jelly also illustrates the relatively disfluent, repetitious, and sometimes idiosyncratic linguistic constructions that were typical of 6- to 10-year-old children’s speech to the computer, which clearly would be difficult for current speech technology to process. In fact, figure X.6 clarifies that 5 percent of children’s spoken language during these interactions involved disfluencies or extreme mispronunciations, invented or misapplied words, and ill-formed syntactic constructions. Given that children’s average MLU was five words long, this means that one in four utterances during this application contained hard-to-process language. These data underscore the need for collection of child speech corpora and the development of language models that represent these unique features of children’s spontaneous interactive speech. The aim of this approach is to leverage more robust spoken language systems by using models of children’s preexisting language patterns, rather than engaging in futile attempts to retrain children’s entrenched and immature speech patterns.

This study confirms that children’s disfluency rates are substantially and consistently higher during interpersonal than human-computer communication, even though the latter interaction is with an animated and embodied “persona.” In fact, the disfluency rate during children’s interpersonal communication averaged three times higher, and all ten children had higher disfluency rates when speaking with a real person, as illustrated in figure X.4. This difference is consistent with the 2.5 to 3 times higher disfluency rate that is typical of adult speech during human-human communication (Oviatt 1995, 1997).

One implication of these findings is that 6- to 10-year-old children do distinguish animated from real human partners, even though at times they may seem completely absorbed with their “new friends.” Children’s speech is not only distinctly different to a computer partner, like adult speech, but also clearer or more hyperarticulated (Oviatt, MacEachern, and Levow 1998). This may indicate that children view their computer partner as a kind of “at-risk” listener, with whom they anticipate needing to work harder to make themselves understood. Children’s higher rate of repetitive speech to the computer also provides evidence of linguistic simplification, compared with language directed to a human adult listener.

In some noteworthy respects, children’s speech to an animated computer partner was significantly more difficult to process than typical adult speech to a computer. After controlling for utterance length, their disfluency rates averaged 64 percent higher than comparable adult speech during a spatially oriented task. As far as we know, analyses have not yet been computed for adult disfluency
rates while conversing specifically with animated characters. Such data will be needed to interpret clearly whether children’s disfluency rates are elevated above those of adults in general, or whether disfluency rates may be higher for both children and adults when talking to animated characters than to other types of spatial display.

Children’s spontaneous spoken disfluencies also were confirmed to be strikingly sensitive to utterance length. Of all the variance in spoken disfluencies in the present data, 88 percent was predictable simply by knowing an utterance’s specific length. This rate is even higher than the 79 percent strength of association found previously for adult speech. These child speech data replicate and extend previous adult findings during verbal, numeric, and spatial tasks (Oviatt 1995, 1997). However, figure X.5 clarifies that the slope of this function is steeper in children’s speech than in adults’. Basically, higher disfluency rates are precipitated at lower utterance lengths in children’s speech. In children’s speech, the disfluency rate begins exceeding 3 percent after an utterance length of seven words is reached, whereas a 3 percent disfluency rate is not exceeded in adult speech until twelve words or longer. This indicates that the planning load associated with increased utterance length has a greater impact on children, even within the range of relatively brief to moderate sentence lengths.

X.4 Future Research Directions

Developing effective conversational interfaces, especially ones that incorporate animated characters with whom users can interact, will require research on many key issues that still are poorly understood. Among other things, research is needed on the basic functionality and best uses of animated characters, and what their central design elements should be. There is great interest in the apparent power of new conversational interfaces to engage and motivate users, perhaps especially child users. As a result, further research needs to explore much more thoroughly the educational role that such technology could play, as well as how it could best be designed for such purposes.

More careful consideration also is needed of the entire interactive input-output cycle between human and computer. For example, what is the impact of character animation, TTS, and other audio output on subsequent user input? In particular, how is users’ speech influenced by animated displays, in terms of the synchronization, complexity, and processability of spoken language that is elicited? And how might the future design of animated characters be used to assist in simplifying or managing the most difficult features of users’ spoken language?
The design of interactive technologies for children, especially ones that have recognition rates matching those of adult users, will require a substantial program of research analyzing children’s spoken language to conversational interfaces. This research will need to include a comprehensive assessment of the acoustic-prosodic, syntactic, semantic, and discourse-level properties of children’s speech. In particular, future analyses should continue investigating features of children’s speech other than disfluencies that are especially difficult for current technology to process. This research ideally should be conducted in natural field contexts in which noise levels, collaborative interaction and other factors are expected to elevate the rate of hard-to-process speech, such as disfluencies, and to reduce system recognition rates. To accomplish the above goals, large corpus collection efforts will be needed on children of different ages while engaged in different types of interaction. In addition, new language models will be needed that represent the central and sometimes idiosyncratic features of children’s spoken language. The new Mobile Simulation Environment, in conjunction with the I SEE! application and others to be developed in the future, currently is being used to address these and related research questions.

Acknowledgments

Special thanks to Intel and NSF grant IRI-9530666 for providing funding and equipment donations in support of this research. Thanks to William Fiesterman for designing our dolphin character and eyeball animations, and to the National Geographic Society and its photographers for permission to use videotape footage from “Jewels of the Caribbean Sea.” Thanks also to Jeff Schilling of OGI’s television productions for video editing assistance; to CHCC’s Josh Clow, Thomas Marsh, Craig Minor, and Joel Pothering for assistance with implementing and testing the new simulation environment; to Jennifer Bruns for assistance with data collection and analysis; and to Phil Cohen for helpful discussions. Finally, special thanks to the children who played with the system, served as design advisors, and inspired this work in the first place.

Notes

1. An “unconstrained” interface refers here to one involving a user-initiated interaction, rather than being system prompted or directed. With respect to dialogue initiative, then, these data are parallel to the present child data, in which queries were initiated by the user. This is important because the degree of user versus system initiative is known to influence spoken disfluency rates (Oviatt 1995).
References


<table>
<thead>
<tr>
<th>Type of spoken interaction</th>
<th>Disfluency rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-human speech</td>
<td></td>
</tr>
<tr>
<td>Two-person telephone call</td>
<td>8.83</td>
</tr>
<tr>
<td>Three-person interpreted telephone call</td>
<td>6.25</td>
</tr>
<tr>
<td>Two-person face-to-face dialogue</td>
<td>5.50</td>
</tr>
<tr>
<td>One-person noninteractive monologue</td>
<td>3.60</td>
</tr>
<tr>
<td>Human-computer speech</td>
<td></td>
</tr>
<tr>
<td>Unconstrained computer interaction</td>
<td>1.74/ 1.87/ 2.14</td>
</tr>
<tr>
<td>Structured computer interaction</td>
<td>0.78/ 0.87/ 1.70</td>
</tr>
</tbody>
</table>

Table X.1 Spoken disfluency rates per 100 words for different types of human-human and simulated human-computer interaction. Human-computer disfluency rates from left to right represent verbal temporal content, computational-numeric content, and visual-spatial content.
Figure 2

System Architecture

AAA Facilitator/Blackboard

Video Agent (ACL, COM)  Microsoft Agent (ACL, COM)

Inker Agent via DCOM

Subject's Display  Wizard's Display
Figure 3

Can this animal change colors?

Stop video
Figure Captions

Figure 1. Child interacting multimodally with I SEE! on hand-held device at Maine Science Center (top) while programmer assistant simulates system and collects user data remotely over a wireless network (bottom).

Figure 2. Mobile Simulation Environment’s system architecture, with user’s display on the left, and linked programmer assistant’s display and control palette on the right.

Figure 3. I SEE! User interface.

Figure 4. Rate of spoken disfluencies during children’s human-computer and interpersonal communication for each of the ten study participants.

Figure 5. Linear regression model summarizing increasing rate of spoken disfluencies per 100 words as a function of utterance length during a visual-spatial task, comparing 6-to-10-year-old children with adults.

Figure 6. Cumulative percentage of children’s human‐computer versus interpersonal speech that contains hard-to-process disfluencies and idiosyncratic linguistic constructions.