Creating a Spoken Impact: Encouraging Vocalization through Audio Visual Feedback in Children with ASD

Joshua Hailpern
Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61802 USA
1-217-333-3328
jhailpe2@cs.uiuc.edu

Karrie Karahalios
Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61802 USA
1-217-265-6841
kkarahal@cs.uiuc.edu

Jim Halle
Special Education
University of Illinois
1310 South Sixth St.
Champaign, IL 61820 USA
1-217-244-3557
halle@uiuc.edu

ABSTRACT

One hallmark difficulty of children with Autism Spectrum Disorder (ASD) centers on communication and speech. Research into computer visualizations of voice has been shown to influence conversational patterns and allow users to reflect upon their speech. In this paper, we present the Spoken Impact Project (SIP), an effort to examine the effect of audio and visual feedback on vocalizations in low-functioning children with ASD by providing them with additional means of understanding and exploring their voice. This research spans over 12 months, including the creation of multiple software packages and detailed analysis of more than 20 hours of experimental video. SIP demonstrates the potential of computer generated audio and visual feedback to encourage vocalizations of children with ASD.

Author Keywords

Accessibility, Visualization, Autism, Children, Speech, Vocalization

ACM Classification Keywords

H5.2 [Information Interfaces and Presentation]: Screen design, Voice I/O. K4.2 [Social Issues]: Assistive technologies for persons with disabilities

INTRODUCTION

As a child develops, acquisition of speech and language typically progresses with little or no explicit effort from parents, family, or doctors. Developmental disorders, such as Autism Spectrum Disorder (ASD), can significantly disrupt the natural development of social behaviors, such as spoken communication. Since language is "a unique characteristic of human behavior… [that] contributes in a major way to human thought and reasoning" [27], the communication deficits of children with ASD are likely to have detrimental effects on multiple aspects of their lives. The impact of this disability as well as its prevalence, estimated by the Center of Disease Control and Prevention (CDC) as 1 in 150 children [10], highlight the need for effective methods to facilitate the development of communication, including speech.

This paper presents SIP, the *Spoken Impact Project,* which aims to explore a new area of HCI: using real-time audio/visual feedback to facilitate speech-like vocalizations in low-functioning children with ASD. This work is grounded in HCI and behavioral science literature. We believe computer-generated feedback, based upon a child's vocalizations, can influence the vocalizations of children with ASD for communicative purposes by providing them with additional means of accessing information regarding parameters of their voice (e.g., pitch, loudness, duration).

We first outline the foundations of SIP's design. We then describe the four areas of our research: Software Design, Within-Subject Experimentation, Data Gathering, and Data Analysis. Beyond the results of the experiment, the main contributions of this work are the demonstration of a new approach to ASD research (within the context of HCI research) and an initial understanding of how the SIP model could be further explored by the HCI community.

LITERATURE REVIEW

Kanner's 1943 description [19] of 11 children with ASD documented this disorder in the scientific community. In the past 60 years, scientists and therapists have strived to better understand ASD and provide treatments to mitigate its many communicative and social difficulties. The ASD population is not a homogenous group. Many of the characteristic difficulties and developmental delays revolve around communication, empathy, social functioning, and expression. The Autism Society of America describes ASD as "insistence on sameness… preference to being alone… spinning objects [and] obsessive attachments to objects"[2]. While some children have limited impairment, those who encounter greater challenges with social and communicative skills are considered low functioning.

Communication Treatments

In the 1960s, Ivar Lovaas' adopted the pioneering approach of "applied behavior analysis" to help teach communication and social skills to children with ASD. The treatment focuses on extrinsic rewards (e.g., food or toys) for encouraging targeted behavior [27]. Over time, rewards are slowly faded/removed resulting in more naturalistic behavior. Other forms of communication treatment [13, 24, 34, 44]

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

have been used to help develop social and communicative skills in children with ASD.

While the merits of this treatment have been documented for 40 years, this form of intervention is financially expensive and labor-intensive. Furthermore, frequent sessions requiring sustained attention and intense human-to-human contact can be anxiety producing [6]. This anxiety, that characterizes many children with ASD, along with their human detachment [6, 19] causes difficulty for practitioners and children. Generalization of social communication skills is another programming challenge. Given the documented impact of these forms of treatment, technology has the potential to augment and/or increase this impact while reducing the burden on families and children.

HCI and ASD Research

Since the 1990s, the HCI community has examined how computers can aid in the diagnosis of ASD [17, 22, 23]. In addition, HCI has studied audio perception [38] and teaching human-to-human interaction to high-functioning children with ASD [21, 25, 31, 43]. Elements of play have also been studied demonstrating that technology/computers can reduce the apprehension associated with human-to-human interaction [25, 29, 35]. Other HCI research [7, 18] and technology-based behavioral science research [1, 5, 40], with individuals other than those with ASD, have illustrated the use of computer solutions in the context of speech and communication therapy.

With this work, we explore methods and technology that can facilitate speech and vocalization processes of children with communication skill deficits. Specifically, we employed contingent visual and auditory feedback (a) to motivate and reward vocalization and (b) to provide information about the acoustic properties of vocalizations. Due to the current limitations of speech recognition software [33, 41], forms of speech detection are limited, especially for individuals with poor diction. Hence, our technological solutions must be designed to aid and supplement practitioners and researchers rather than replace them.

SCOPE AND MOTIVATION

SIP explores a new area of HCI research focusing on the use of contingent audio and/or visual feedback to encourage sound production in low-functioning children with ASD. Without the development of techniques to encourage speech/vocalization, a diagnosis of ASD can have far reaching negative implications for a child's social, developmental and educational life.

Building on prior work, our focus on computer visualization in this population is unique. Most HCI visualizations research has focused on neurologically typical individuals [39]. ASD treatment research in HCI has targeted higher functioning children with ASD [42], failing to address the needs of non-verbal/low-functioning children with ASD. These higher-functioning individuals tend to be more capable of articulating their needs and opinions during evaluation as well as understanding tasks they may be asked to complete. Although the literature in the behavioral sciences has explored this demographic, many existing practices use traditional alternatives such as PECS [8], mirrors and echo chambers [28] or invasive procedures, such as electropalatography [9]. Our research begins with the basic question: can real-time visual/audio feedback positively impact sound production in low-functioning children with ASD?

While there is controversy about whether high-functioning children with ASD should be pressured to communicate vocally, this concern is not applicable to this vein of research. These children do not possess a linguistic system (e.g., typing, signing or speaking). Teaching functional communication is essential, although the method should vary according to individual preference and capabilities.

RESEARCH QUESTIONS

We posed the following research questions about the effects of contingent audio and/or visual feedback on low functioning children with ASD.

R1: Will at least one form of real time computer-generated feedback positively impact the rate of spontaneous speech-like vocalizations?

R1 is the primary question of SIP: testing the impact of computer-generated feedback. R1 builds upon the success of traditional alternatives (e.g., image cards [8], mirrors [28]) and other related work. The remaining research questions examine modes of feedback, and their implications on rate of spontaneous speech-like vocalization. R2-R5 are derived from research into the profiles of children with ASD [26, 37] that concluded that these individuals prefer visual feedback [3, 12, 30, 32]. The responses to R2-R5 will directly impact future systems and the extent to which individualization is needed.

R2: Will all forms of feedback positively impact the rate of spontaneous speech-like vocalizations?

R3: Will subjects increase the rate of their spontaneous speech-like vocalizations during conditions with *visual-only* feedback, *audio-only* feedback and/or *mixed* feedback?

R3a: If there is a modality that demonstrates significance (R3), or $p < 0.1$, which specific form of feedback in that modality significantly and positively impacts the rate of spontaneous speech-like vocalizations?

The quantitatively driven investigation of R3 may hide the impact of a specific form of feedback; if one form of feedback in a modality is significant, but fails to adjust the results in R3, it will not be analyzed in R3a. Therefore R4;

R4: By testing feedback conditions that were qualitatively favored by subjects (assessed during the experiment and via video), will we uncover forms of feedback that positively impact the rate of spontaneous speech-like vocalizations?

R5: Is there a modality of feedback whose variations indicate (R3, R3a, and R4) the child's rate of spontaneous speech-like vocalization are positively impacted.

EXPERIMENTAL SETTING

This paper is the culmination of more than 12 months of research. The process consists of four main areas: Software

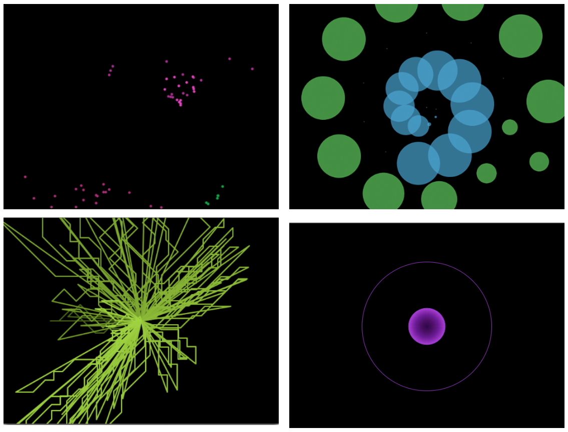

Figure 1. Examples of visualizations used in SIPS.

Design, Within-Subject Experimentation, Data Gathering, and Data Analysis.

Software Design

During three months (Summer 2007), the researchers designed the *Spoken Impact Project Software (SIPS)* package in Java using the Processing Library [11]. SIPS generates audio and visual feedback directly related to the amount of external noise detected by the system. For example, a circle on the screen could change in diameter, as sound, particularly voice, grows louder. An "echo", like that heard in a stairwell, is an example of audio feedback. Distortions could be applied to change the perception of the sound returned to the subject.

SIPS visual feedback (Figure 1) consists of one of three possible types of graphical objects: circular/spherical, lines, or found images (e.g., picture of cartoon character). These objects can be presented in one of four types of motion metaphors: **(1) Falling**–objects move from one portion of the screen downward, as if drawn by gravity. This includes particle effects like water from a shower head or fireworks (Figure 1, top left); **(2) Spinning**–objects move in a circular or spiral pattern (Figure 1, top right)**; (3) Flashing**–objects appear and disappear quickly (Figure 1, bottom left); **(4) Stationary**–objects stay in a fixed location (Figure 1, bottom right).

The falling and spinning metaphors were selected to leverage stimuli that garner interest from children with ASD [3, 12, 32]. Flashing feedback was investigated due to its high energy which often appeals to neurologically typical children. Stationary objects were explored to focus on change in an object (size, color, etc.) rather than object motion. Among the four categories, approximately 12 unique motion/pattern combinations were created; most can function with any type of object (circle, found image, etc).

SIPS provided two categories of audio feedback based on the sound produced.

- 1-to-1: sound produced by the interface was directly related to sound produced by the subject (e.g., echo, or pitch-shifted version of the subject's voice). There was a slight delay between the source sound and feedback, but both input and output occured simultaneously.
- Reward: computer sound was produced upon completion of subject's sound. Duration of reward sound was related to duration of sound produced (longer sound made by subject resulted in longer reward). Sound could be music or found-audio (e.g., from movie or TV show).

There were five forms of audio feedback available that could be mixed with any visual feedback permutation.

Within-Subject Experimentation

Our subjects demonstrated limited response to requests or instructions to perform tasks due to the severity of their ASD. Therefore, engaging subjects in the same activity across trials and sessions was not a viable option. We relied on the visual/auditory feedback to be sufficiently engaging to promote spontaneous speech-like vocalizations. The feedback presented and tested was varied across children to enable an exploration of R3 and R3a. As a result, each child's performance served as his or her own baseline/control for comparison across conditions. Given the number of subjects participating and the questions generated, a within-subject design was selected. The analyses were conducted using a baseline created by each child and comparing that baseline to each of the computerized feedback conditions: visual, auditory or visual/auditory combined.

The within-subject experimental design [20], an adaptation of an alternating treatments design [4], consisted of five non-verbal children (aged 3-8 years) diagnosed with "low-functioning" ASD. Each child enrolled in the study first participated in one to three 30-minute "orientation sessions" which acclimated the child to the study room, researchers, and computer feedback. No data were recorded during these sessions, though initial preferences for feedback type/style were noted. Room configuration was selected based on child preference and is described/labeled in Figure 2.

Each child attended 6 data sessions after completing the orientation period. A data session lasted for approximately 40 minutes and consisted of approximately 8 two-minute trials. During a trial, a researcher exposed the subject to a specific form of feedback (permutations of audio and/or

Figure 2. Clockwise: A) projector screen with open room (with beanbag chair or trampoline) **B)** projector screen with separated work area **C)** large screen computer at desk.

visual). Each trial began with an antecedent demonstration by the researcher (e.g., saying "boo" and pointing to screen to show the visual effect). Subjects then could engage the system in whatever manner they chose.

Feedback permutations were selected based on researcher's subjective calculation of vocalization rate. Order of presentation was randomized across sessions to accommodate for order effects. However, the first trial of each session was a baseline trial with no audio or visual feedback. Although this baseline trial provided a means of comparison for assessing changes in spontaneous speech-like vocalizations due to visual/auditory feedback, we provided no control for order effects related to the presentation of the baseline condition. Baseline was always the first trial.

Data Gathering

Because the subjects did not use spoken language or attend to structured assessments, data collection was limited to observable behavior. We gathered data by analyzing video of each trial through video annotation. To annotate the video and assess coder reliability, we used two tools:

VCode and VData

Examination of existing digital tools for digital video annotation found interfaces to be overly complicated and lacking in reliability measurement functionality. Therefore, we designed/built a suite of applications called *VCode and VData*. A full description of the tools, features, justification and reaction of users is presented in [14]. *VCode* is a tool used by coders to view and annotate digital video. In order to support SIP, *VCode* provides two types of annotations: ranged (having start/end times) and momentary (one time). *VData* is a tool designed to perform agreement calculations and link annotations by two coders back to the original video. *VData* utilizes point-by-point agreement to assess reliability. Point-by-point agreement is calculated by assigning one coder as a *primary coder* and the other as *secondary*. Any mark made by the *primary coders* is considered an opportunity for agreement. Marks made by the *secondary coder* that are the same as the *primary coder* are considered agreements. The percent agreement is calculated by dividing agreements over opportunities. Through *VCode* and *VData,* the data collection and agreement calculation process was simplified.

A^3 Coding Guidelines

A set of dependent variables was developed to quantitatively assess the impact of SIPS. This guide, A^3 (pronounced A-Cubed) or **A**nnotation for **A**SD **A**nalysis, was based on existing literature in HCI and behavioral sciences. A full description of A^3, the 18 variables, developmental process, justifications, coder's guide, reliability data, and reactions from coders is presented in [15].

We focused our analysis on *Spontaneous Speech-Like Vocalizations* (SSLV), one of the dependent variables from A^3. There is a clear and important distinction between those vocalizations that are spontaneous and those that are imitative. This is critical when assessing children with special needs [16].

> *Spontaneous Speech-Like Vocalizations* (SSLV)– sounds produced by the subject that can be phonetically transcribed (sounds that could be useful in oral communications) and are not being imitated.

Unlike imitated vocalizations (echolalia), SSLVs are more indicative of vocalizations that may be used for meaningful speech because they rely on longer-term storage and retrieval of linguistic information [15].

Data Collection

Over a six-month period, 1200 minutes of video were annotated (20-40 minutes/ to annotate one minute of video). One random video from each session was tested for reliability using point-by-point agreement calculations[1]. Inter-rater agreement (IRA) across all 18 variables was 88%. For this paper, we used SSLV as the dependent variable, whose IRA for occurrence was 85%, and durational agreement was 93% (because SSLV in A^3 is not a durational variable, durational values were gathered by filtering *Speech-Like Vocalizations* (durational) for those that were *Spontaneous*.

Dependent and Independent Variables

Our within-subject experiment analyzed the dependent variable *Spontaneous Speech-Like Vocalization* (SSLV). The independent variables were the various permutations of visual and auditory feedback. These analyses permitted comparisons between the mode of feedback (visual, auditory, and mixed) as well as the different types of feedback within modes (12 visual and 5 auditory forms).

Data Analysis

Each subject was analyzed independently. Due to the varying lengths of each trial, a comparison of the number of occurrences of SSLV would be weighted towards longer sessions. To mitigate this effect, we analyzed a normalized rate of SSLV (occurrences in trial divided by trial duration) to arrive at a rate. Wilcoxon rank-sum and Kruskal-Wallis tests were used to compare SSLV rates in response to different types of feedback. The Wilcoxon rank-sum test is a non-parametric alternative to the paired T-test. The Kruskal-Wallis test is a non-parametric alternative to a one-way analysis of variance (ANOVA). These tests were well suited for these data where distributions were not normal and where numbers were small because they do not make any distributional assumptions. All tests used a two-tailed alpha with a $p<0.05$ denoting statistical significance.

R1 Analysis

R1 examines if there is at least one form of computer generated feedback that will positively impact a subject's rate of SSLV. If at least one condition (R2-R4) shows that feedback has a positive impact on rate of SSLV, we can conclude R1 is true for that subject.

---

[1] Cohen's Kappa [20] is not applicable to use for agreement in this case since this variable is coded on an "infinite", or continuous scale (the mark locations were not explicitly broken up into discrete sections, and thus, chance agreement is not applicable).

	Age	Diagnosis	Room Setup	Any Feedback	Visual Only	Audio Only	Mixed Feedback
Oliver	5	ASD	C	*0.065 [-1.85]*	*0.386[-0.87]*	*0.063[-1.86]*	*0.058[-1.89]*
Frank	8	ASD + Downs	C	*0.024 [-2.26]*	*0.556[0.59]*	*0.011[-2.56]*	*0.006 [-2.71]*
Larry	4	ASD + Downs	C	*0.850 [-0.19]*	*0.796 [0.26]*	*0.805 [0.25]*	*0.650[-0.45]*
Diana*	4	ASD	B	*0.789 [-0.27]*	*0.016 [-2.41]*	*not used*	*0.470 [0.72]*
Brian	3	ASD	A	*0.834 [0.21]*	*0.766 [0.30]*	*not used*	*0.796 [-0.26]*

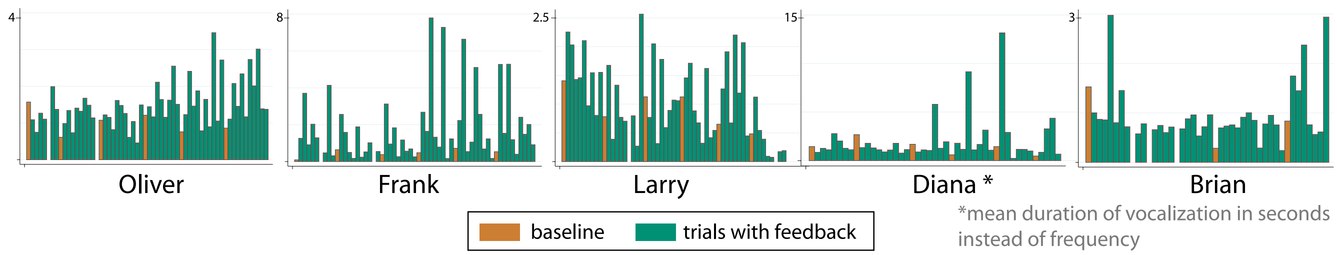

Figure 3. Demographics and Rate of SSLV: Wilcoxon Rank-Sum Test from R2 and R3 Analysis. High level graphical comparison of Rate of SSLV per 10 seconds across all trials for all subjects.

R2 Analysis

R2 indicates, in general, that all forms of feedback (regardless of mode/type) increase rate of SSLV. An examination of R2 for each subject is determined by comparing the rate of SSLV at baseline to rate across all types of feedback using the Wilcoxon rank-sum test.

R3 Analysis

R3 addresses whether all forms of feedback in a specific modality positively impact SSLV. An examination of R3 for each subject is conducted by performing a Wilcoxon rank-sum test comparing rate of SSLV at baseline with rate of SSLV in groups *audio only*, *video only*, and *mixed feedback*. Results from R3 can be **Video** (*video only* significant $p<0.05$), **Audio** (*audio only* $p<0.05$), **Mixed** (*mixed feedback* only $p<0.05$) or some permutation of the three. If none has a significant p value, R3 is considered **Neither**, indicating that no modality increased the rate of SSLV (all $p>=0.05$).

R3a Analysis

R3a examines if there is a specific type of feedback that increased the rate of SSLV in a modality that approached significance. Using the result from R3, we parsed specific forms/combinations of feedback within those statistically significant modalities (*visual, auditory, mixed*). Trials within the specific modality are divided into subcategories based on specific forms of feedback and tested against baseline using the Wilcoxon rank-sum test. R3a required that p values for R3 either approach or reach statistical significance.

R4 Analysis

We used qualitative observations gleaned from researchers and video to guide this analysis. Here we could employ overlooked forms of feedback that *appeared* to increase the rate of SSLV. Using the Wilcoxon rank-sum test, we compared baseline with conditions that were qualitatively observed to increase SSLV rates.

If significance was not found, the Kruskall-Wallis test was used to determine if differences existed in SSLV rates across feedback type, excluding baseline comparisons. This additional analysis allows us to compare the impact of one type of feedback against all others.

R5 Analysis

In order to categorize the types of feedback which illicit an increase in SSLV rate, we extracted the mode of feedback found to have the most impact in R3, R3a and R4. This synthesis of results provides a better understanding of what specific modes of feedback are engaging for each child.

RESULTS

To protect the privacy of our subjects, we have changed their names; gender status was maintained. All five of the subjects' spoken language developmental benchmarks [36] were in the first phase (Preverbal Communication), equating closely to the development of a neurologically typical 6-12 month old.

Subject 1: Oliver

Oliver's Results

Initial analysis of Oliver's data (Figure 3) demonstrated borderline significance comparing baseline to all feedback trials (R2). Further, the *audio only* and *mixed feedback* conditions (R4) approached significance. Due to a trend towards significance in the two conditions involving audio, we compared the rates of SSLVs at baseline with *any* condition containing audio feedback (both with and without visual feedback). There was a statistically significant difference between conditions containing *any* audio feedback and those containing no audio (*p=0.045 [-2.00]*). We concluded

	Found Audio	Echo
Audio Without Visual	*0.200 [-1.28]*	*0.045 [-2.00]*
Audio With Visual	*0.082 [-1.74]*	*0.073 [-1.79]*
Any Condition	*0.076 [-1.77]*	*0.04 2[-2.03]*

Table 1. Comparison of Oliver's audio feedback

Audio Feedback	p value with visual feedback	p value without visual feedback
Any Found Audio	*0.010 [-2.57]*	*0.011 [-2.56]*
Child's Cartoon Found Audio	*0.003 [-2.98]*	*0.011 [-2.56]*
Echo	*0.005 [-2.80]*	*No data*

Table 2. Comparison of Frank's audio feedback

that audio feedback may have played a role in increasing the rate of Oliver's SSLV's (R5).

Since audio appeared to increase the rate of Oliver's SSLV, we explored the impact of different types of audio feedback in combination with visual feedback. Table 1 suggests that echo feedback encouraged SSLV, while visual feedback did not appear to have significant impact on SSLV rate (R3a). We qualitatively observed that Oliver reacted positively to audio from a popular cartoon show. Our data confirmed this by approaching statistical significance (*p=0.082 [-1.74]*) (R4).

From this analysis, we conclude that audio feedback was associated with higher rates of SSLV for Oliver. Specifically, SSLV rates increased in conditions with echoing audio feedback (R1).

Subject 2: Frank

Frank's Results

Initial analysis of Frank's data (Figure 3) showed a significant difference between baseline SSLV rates and rates produced by all feedback conditions (R2). We found a statistically significant difference in rates of SSLV with *audio only* and *mixed feedback* (R3). Due to significance in both conditions with audio, we compared rate of baseline SSLV with *any* condition with audio feedback. There was a highly significant association (*p = 0.004 [-2.84]*) (R5).

Given the robust effect of audio feedback, we compared Frank's responsiveness to audio feedback with and without visual feedback (Table 2). Audio feedback was categorized as "found audio" and "echo". Based on our qualitative observations, we isolated and analyzed trials where audio feedback from a specific child's cartoon was present. Frank demonstrated the most significant increase in rate of SSLV over baseline when audio from the cartoon was present (R3a, R4). For Frank, visual feedback had a positive impact on the SSLV rates when audio was also present.

Finally, we examined all conditions with audio feedback into those with specific types of visual feedback to assess the impact of specific visual feedback types on the rate of SSLV production. Based on qualitative observations, we analyzed trials where a visual image from a specific cartoon was present. Frank demonstrated increased rate of SSLV over baseline for all visual feedback in addition to audio - except for Spinning Spiral of Dots and Random Dots (Table 3), with the clearest effect in Firework-Like (R3a).

Form of Visual Feedback in Addition to Audio	P Value
No Visual Feedback	0.011 [-2.56]
Cartoon Image	0.046 [-2.00]
Firework-like	0.004 [-2.86]
Spinning Spiral of Dots	0.160 [-1.41]
Fast Flash	0.010 [-2.58]
Line Circle	0.032 [-2.14]
Random Dots	0.134 [-1.50]
Shower	0.046 [-2.00]

Table 3. Frank: Form of visual feedback with any audio

From this analysis, we concluded that Frank had higher rates of SSLV in conditions with audio feedback and with audio and visual feedback combined (R1 , R5). Specifically, he appeared to produce higher rates of SSLV in response to audio and visuals from a specific cartoon. Interestingly, his mother stated that Frank did not watch this cartoon show.

Subject: Larry

Larry's Resuxlts

Initial examination of Larry's data (Figure 3) failed to reach statistical significance for the R2 and R3 analyses. While statistical tests did not reach statistical significance, qualitative observations from researchers and study video, in conjunction with graphical representation of the data (Figure 4), led us to believe that feedback did have an impact on rate of SSLV, specifically conditions with echoing audio feedback. Qualitatively, researchers observed a higher degree of attention and SSLV, during conditions with echo/reverb feedback.

Comparing conditions with echoing feedback with baseline produced a lower p-value than other analysis (*p=0.243[-1.16]*), yet it did not reach $p<0.05$. To examine the impact of echoing feedback, we repeated our analysis

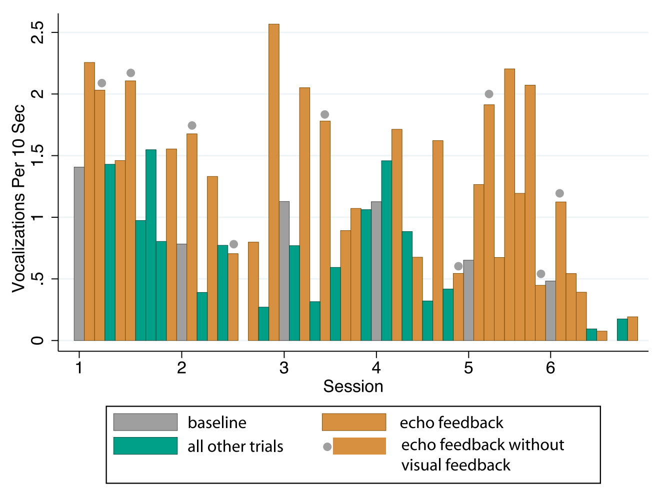

Figure 4. Larry's SSLV rate, by session and trial.

	0	1	2
0	*X*	*X*	*X*
1	*0.284 [-1.07]*	*X*	*X*
2	*0.055 [1.91]*	*0.034 [2.13]*	*X*
3	*0.410 [0.83]*	*0.023 [2.27]*	*0.396 [-0.85]*

Table 4. Larry's comparative conditions (Row vs. Col). 0=baseline; 1=Any Condition with ECHO; 2=Audio only; 3= Mixed + Visual Only

across test conditions. We performed a Wilcoxon rank-sum test to compare conditions using echoing feedback with visual feedback to conditions with only echoing feedback and no visual feedback. Given $p=0.970$, we concluded that there was no significant difference in SSLV between echoing conditions with and without visual feedback.

To compare the impact of echoing feedback on SSLV rates with varying types of feedback, we used the Kruskal-Wallis test. First, we categorized all of Larry's trials into one of the following 5 conditions; (**1**) baseline, (**2**) any condition *with* echoing feedback, (**3**) audio feedback only (excluding echoing), (**4**) visual feedback only (excluding echoing), (**5**) audio + visual feedback (excluding echoing). The Kruskal-Wallis test resulted in $p=0.060$. To increase statistical power, we combined visual only feedback with the mixed condition since groups had visual presentations[2] (comparative analysis between collapsed groups: Wilcoxon rank-sum $p=1.000[0.00]$). Analysis of these groups found a statistically significant difference ($p=0.030$ by Kruskal-Wallis test). A post hoc pair-wise comparison of each condition, using Wilcoxon rank-sum test (Table 4) was performed. Statistically significant differences were found between the echo condition and audio only(visual + mixed) (p=0.034, $p=0.023$ respectively) (R4).

From this analysis, we concluded that Larry showed preference for echoing audio feedback (R1, R5). However, we believe that with more statistical power, we could make a more conclusive statement.

Subject: Diana

Diana's Results

Initial data analysis for Diana, revealed much higher p values (0.5-0.9) than expected when comparing them to qualitative notes made by the researchers. Confused by these findings, we examined annotations made by video coders and noticed that large strings of Diana's SSLV's were grouped together. A[3] guidelines stated that utterances must be separated by a pause of 2-seconds to be considered independent. However, Diana's pauses ranged from 1-to-1.5 seconds in duration. As a result, phrases of multiple utterances were captured as a single occurrence. To accommodate her shorter pauses, we re-analyzed her data using mean *duration* of SSLVs rather than rate. For this subject, we used average duration as a proxy for rate.

Form of Visual Feedback	P value (without audio)	P value (with audio)
Firework-like	0.136 [-1.49]	0.934 [-0.08]
Spinning Image	0.020 [-2.32]	0.201 [-1.28]
Shower-like	xxx	0.439 [0.78]
Fast Flash	xxx	0.739 [-0.33]
Multiple Circles	0.020 [-2.32]	0.556 [-0.59]
Line Circle	0.617 [0.50]	0.439 [0.78]
Fast Spin	xxx	0.617 [0.50]
Found Imagery	0.003 [-2.97]	0.330 [-0.97]

Table 5. Diana: Forms of Visual Feedback tested, vs. baseline (with and without audio)

Initial analysis of SSLV duration (Figure 3) was significant for visual only conditions (R2, R3). Audio only feedback (conditions with no visual feedback) was not tested, due to lack of interest observed in initial orientation sessions.

To examine impact of visual feedback, we divided the types of visual only feedback and compared the average duration of SSLV's with those produced in the baseline condition (Table 5). The last row in Table 5 is an amalgam of different forms of visual feedback in which abstract colored dots were replaced with one or more found image(s). These data support our qualitative observations that Diana responded only to conditions where images shown were from cartoon shows, and that audio feedback reduced her SSLV duration (R4). Three statistically significant forms of feedback were Spinning Image (a found image from a cartoon spins on axis), Multiple Circles (many dots or found images appear on the screen whose size is based on volume of the child's sound) and any feedback with Found Images (there are overlaps between groups) (R3a, R4).

From this analysis, we concluded that Diana produced more SSLV's (mean duration) with visual feedback compared to baseline and mixed (R1, R5). Specifically, she appeared to show increased engagement with forms of visual feedback that contained a cartoon character (although a specific preference did not appear). Diana's mother reported that she watched movies/TV-shows with these characters.

Subject: Brian

Brian's Results

Brian was the most difficult subject for us to qualitatively discern a particular pattern or "taste" for feedback. This was supported by extremely high p-values for all coarse tests conducted (Figure 3). During three sessions, we inadvertently failed to run a baseline, reducing the number of comparison points to three instead of six. This reduced statistical power. While Wilcoxon rank-sum statistics approached significance for one particular form of visualization in which a cartoon character spun in a circle centered on screen, it failed to reach significance.

[2] Collapsing two groups increases the number of data points in the resulting group, thus increasing the statistical power during comparison.

From these analyses, we could not conclude that Brian had a significant response to any type of feedback (either in comparison to baseline or against each other) (R1-R5).

DISCUSSION

After a thorough examination of the quantitative data collected, we have summarized the findings in relation to our 5 questions in Table 6. Given the constraints of these single subject analyses and our small sample size, we present our results.

R1

In 4 of the 5 subjects, we found that at least one type of feedback produced an increased rate of SSLV. We were unable to show that any form or modality of feedback, when compared to baseline, significantly increased the rate of SSLV's for Larry and Brian. This may be, in part, due to the small number of data points collected and high degree of ASD. We were, however, able to demonstrate that echoing audio feedback produced a significant difference in rate of SSLV's when compared with all other forms of feedback for Larry. Overall, we concluded that particular modes/ types of feedback may encourage SSLV in children with ASD.

R2

Only one of five subjects found all forms of feedback, regardless of mode or type, to have a positive impact on rate of SSLV. This finding suggests that not all forms of computer feedback are equally influential for all children.

R3 & R5

It is commonly believed that individuals with ASD are more responsive to visual feedback than auditory [3, 12, 32]. However, we had two subjects who responded primarily to auditory feedback (Oliver and Larry). One preferred a mixed condition (Frank). One responded to visual only (Diana). And one subject (Brian) did not show any significant response to any form of feedback. When viewed from a global level, 3 of 5 subjects responded to audio feedback, and 2 of 5 responded to visual feedback (Table 6). This suggests that further exploration of feedback in both the visual *and* audio modalities is essential. This finding is of particular note in that it is in contrast to other work.

R3a & R4

Although some subjects had a larger range of forms of feedback that resulted in increased rate of SSLV than others, 4 of 5 subjects did have one particular condition that out-performed the others. The specific results, in conjunction with varied modes of feedback that resulted from R3 analysis, indicate that visualizations, and any potential therapeutic application, will likely need to be tailored to individual subjects. The degree of customization is unknown due to small sample size. We can proceed, however, knowing that individual interests/preferences must be taken into consideration. This work illustrates the varied types of audio/visual feedback that garnered the increase in SSLV.

Parental Response

In addition to data from subjects during the sessions, we solicited anonymous parental response in the form of a written questionnaire. Feedback from parents was positive and encouraging. Parents responded with high praise for our technique, and asked for similar procedures to be implemented in their own homes. One mother stated,

> *My child's reaction is one of excitement and looking forward to see what was next to come. Applause on your study. You may be onto something here.*

Another mother stated her child's reaction,

> *Since my son is fairly severely affected by autism, he stays in his "own little world" quite a bit. So the fact that he showed interest in and seemed to enjoy some of the visuals and sounds is quite a positive thing. Thank you.*

	R1	R2	R3	R4	R5
Oliver	P	X	X	P	A
Frank	P	P	A + M	P	A + M
Larry	P	X	X	P	A
Diana	P	X	V	P	V
Brian	X	X	X	X	X

Table 6. Results by subject.
P = Positive, X = Negative, A = Audio, V = Visual, M = Mixed

FOLLOW UP STUDY

Researchers qualitatively noted that Frank's response was exceptional, both in terms of his reaction to the computer feedback and his eagerness to participate. Noting this, researchers constructed a Wizard-of-Oz system, based on SIP, geared towards teaching specific skills. The model followed a common form of behavioral therapy [27]: Prompt a word – wait for a response – reward if correct or repeat if incorrect. We replaced the computer voice recognition with a researcher to test the concept.

This system aurally prompted subjects with a word in the form of the phrase "Say [word]." Once the prompt was completed, the computer provided visual feedback (spinning spiral of dots) and audio feedback (echo). Immediate feedback provided the subject with an instantaneous reaction to their sounds, for both visual and auditory reinterpretation. If Frank did not repeat the sound, or the repeated sound was not "close enough," the researcher directed the system to re-prompt. If Frank's response was "close enough," the researcher directed the system to provide an auditory and visual reward.

With parental permission, we conducted 2 sessions using this system. The first consisted of 10 words, which had been previously used by Frank (according to his mother). Initially, Frank played with the system (similar to SIP sessions). After 15 minutes, he began repeating words upon the request of the system. At the end of the 30-minute period, Frank imitated every prompted word.

During the second session, we used 6 words his mother stated that he had not spoken before, in addition to 4 words from the previous session. We asked Frank's mother to provide us with words she hoped he would learn, but has not used to date. Frank readily played the Prompt-Repeat game and attempted to imitate the new words. Although articulation was often unclear, he made a concerted effort to repeat all 10 words, including the 6 new ones. Of particular note, Frank had been highly resistant in the past to this form of vocal imitation language therapy.

While this follow-up study consisted of a sample of n=1, and was conducted using Wizard-of-Oz, it does suggest the possible implications of using a SIP-like system to encourage and teach specific vocal skill imitation or possibly acquisition.

FUTURE WORK

Given our encouraging results, there are many exciting areas of future work. One of the most immediate directions might be adaptive feedback selection. Previously, researchers subjectively assessed which visualizations and forms of audio feedback were engaging to subjects. Future work might examine if a system could adaptively change forms of feedback in concert with the subject's response via machine learning. This would not only ease the job of clinicians and researchers, but as preferences change and subjects satiate, such a system would accommodate and adapt to these changing preferences.

We see the potential to test our approach with other populations or other target behaviors. One unanswered question is how to identify a method for teaching specific vocal skills, such as words in context, syllables, etc. Another opportunity would be to explore the delivery of a SIP appliance. The investigation of a toy-like device could provide therapeutic play at home, as well as at the practitioner's office.

LIMITATIONS

The children participating were diagnosed with autism and had significant intellectual disabilities. Their attention to tasks was limited. Sometimes the subjects would appear highly engaged with a type of feedback, while other types proved completely unengaging. This often resulted in trial sessions of extremely short duration, as subjects would get up and move away from the computer. Duration of the trials had high variance, and reduction in observation time may have reduced statistical power of this study and the capacity for statistical tests to reach significance. We may not have fully appreciated the positive effects of *SIPS* in this small study. However, we were able to observe numerous types of feedback that garnered significant changes in SSLV rates.

With the small scale of this first study, we cannot conclude that audio/visual feedback will increase SSLV for every child with ASD. However, based on our 5 within subject analyses, we believe our results are promising.

We also wish to highlight that there is a leap between producing SSLV and real-world communication. Our current study focused specifically on encouraging a behavior that is a precursor for functional communication. This work, in conjunction with the findings from the Follow Up study, lay the ground work for future exploration of this area of research.

CONCLUSIONS

Given the results from the SIP study, we believe that audio and/or visual feedback can be used to encourage spontaneous speech-like vocalizations in low-functioning children with ASD. In addition, SIP demonstrated that both visual *and* auditory feedback can impact spontaneous speech-like vocalization. This suggests that further exploration of feedback in both modalities is essential. This finding is of particular note in that it is in contrast to other existing work.

SIP also suggests that low-functioning children with ASD may have distinct and varied preferences for types/styles of feedback. As a result, individual customization may be necessary in future efforts. Although the range of variation necessary is unknown, the final solution might include a suite or menu of feedback styles that may be selected by the parent, clinician, or child.

Given the promising results of our data, the encouraging messages of parents, and the potential impact demonstrated in the Wizard-of-Oz study, we believe that SIP-styled therapy is an exciting and viable method for encouraging speech and vocalization in low-functioning children with ASD. This research presents initial steps towards exploring the application of audio and visual feedback to encourage speech in low functioning children with autism.

ACKNOWLEDGMENTS

We would like to thank our Collaborators (Laura DeThorne, and Mary-Kelsey Coletto), all participants and their families, our coders (Ashley Sharer, Christine Holloway), NSF (NSF-0643502), Microsoft, our friends, family, and loved ones.

REFERENCES

[1] Adler-Block, M., Bernhardt, B. M., Gick, B. and PBacsfalvi, P. The Use of Ultrasound in Remediation of North American English /r/ in 2 Adolescents. American Journal of Speech-Language Pathology, 16, May (May 2007), 128-139.

[2] Autism Society of America, A. Autism Society of America. http://www.autism-society.org 2007

[3] Baggs, A. In My Language. YouTube, City, 2007.

[4] Barlow, D. H. and Hayes, S. C. Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. Journal of Applied Behavior Analysis, 12, 2 1979), 199-210.

[5] Barry, R. M. EPG from square one: An overview of electropalatography as an aid to therapy. Clinical Linguistics & Phonetics, 3, 1 1989), 81-91.

[6] Baskett, C. B. The effect of live interactive video on the communicative behavior in children with autism. University of North Carolina at Chapel Hill, Chapel Hill, 1996.

[7] Bergstrom, T. and Karahalios, K. Seeing More: Visualizing Audio Cues. In Proceedings of the INTERACT (Rio de Janerio, Brasil, 2007). ACM Press., New York, NY, 2007.

[8] Bondy, A. and Frost, L. The Picture Exchange Communication System. Behavior Modification, 25, 5 2001), 725-744.

[9] Carter, P. and Edwards, S. EPG therapy for children with long-standing speech disorders: predictions and outcomes Clinical Linguistics & Phonetics, 18, 6 (September 2004), 359.

[10] Center for Disease Control and Prevention, C. Autism Information Center, DD, NCBDDD, CDC. http://www.cdc.gov/ncbddd/autism/ Atlanta, 2007

[11] Fry, B. and Reas, C. Processing. 2007

[12] Grandin, T. Thinking in Pictures: And Other Reports from My Life with Autism. Vintage Books, New York, 2006.

[13] Greenspan, S. I. and Wieder, S. Developmental Patterns and Outcomes in Infants and Children with Disorders in Relating and Communicating: A Chart Review of 200 Cases of Children with Autistic Spectrum Diagnoses The Journal of Developmental and Learning Disorders 1, 1 1997).

[14] Hagedorn, J., Hailpern, J. and Karahalios, K. G. VCode and VData: Illustrating a new Framework for Supporting the Video Annotation Workflow. In Proceedings of AVI (Napoli, Italy, 2008). ACM-PRESS, New York, NY, 2008.

[15] Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. A3: A Coding Guideline for HCI+Autism Research using Video Annotation. In Proceedings of ASSETS 2008 (Halifax, Canada, 2008). ACM-PRESS, New York, NY, 2008.

[16] Halle, J. Teaching Language in the Natural Environment: An Analysis of Spontaneity. Journal of the Association for Persons with Severe Handicaps , 12, 1 (Spring 1987), 28-37.

[17] Hayes, G. R., Kientz, J. A., Truong, K. N., White, D. R., Abowd, G. D. and Pering, T. Designing Capture Applications to Support the Education of Children with Autism In Proceedings of Ubicomp 2004 (Nottingham, England), 2004.

[18] IBM. Speech Viewer III. 1997

[19] Kanner, L. Autistic Disturbances of Affective Contact. V.H. Winston, 1943.

[20] Kazdin, A. E. Single-Case Research Designs: Methods for Clinical and Applied Setting. Oxford University Press, USA, 1982.

[21] Kerr, S. J., Neale, H. R. and Cobb, S. V. G. Virtual environments for social skills training: the importance of scaffolding in practice. In Proceedings of ASSETS 2002 (Edinburgh, Scotland, 2002, 2002). ACM Press, New York, NY, 2002.

[22] Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N. and Abowd, G. D. Grow and know: understanding record-keeping needs for tracking the development of young children. In Proceedings of CHI 2007 (San Jose, California, 2007). ACM Press, New York, NY, 2007.

[23] Kientz, J. A., Hayes, G. R., Abowd, G. D. and Grinter, R. E. From the war room to the living room: decision support for home-based therapy teams. In Proceedings of 2006 CSCW (Banff, Alberta, Canada, 2006). ACM Press, New York, NY, 2006.

[24] Koegel, L. K., Koegel, R. L., Harrower, J. K. and Carter, C. M. Pivotal Response Intervention I: Overview of Approach. The Journal of The Association for Persons with Severe Handicaps, 24, 3 (Fall 1999), 174-185.

[25] Lehman, J. F. Toward the use of speech and natural language technology in intervention for a language-disordered population. Proceedings of ASSETS '98 (Marina del Rey, California, United States, 1998). ACM Press, New York, NY, 1998.

[26] Leonard, L. B. Children with Specific Language Impairment. MIT Press, Cambridge, MA, 2000.

[27] Lovaas, I. I. The Autistic Child. John Wiley & Sons, Inc, New York, 1977.

[28] Marshalla, P. Becoming Verbal With Childhood Apraxia: New Insights on Piaget for Today's Therapy. Marshalla Speech and Language, Kirkland, WA, 2001.

[29] Michaud, F. and Théberge-Turmel, C. Mobile robotic toys and autism. Springer, 2002.

[30] Minshew, N. J., Goldstein, G. and Siegel, D. J. Neuropsychologic Functioning in Autism: Profile of a complex information processing disorder. Journal of the International Neuropsychological Society, 3. 1997), 303-316.

[31] Mohamed, A. O., Courboulay, V., Sehaba, K. and Menard, M. Attention analysis in interactive software for children with autism. In Proceedings of the ACM ASSETS 2006 (Portland, Oregon, October 23 - 25, 2006). ACM, 2006.

[32] Mukhopadhyay, T. R. Beyond the Silence: My Life, the World and Autism. National Autistic Society, London, 2000.

[33] Nakagawa, S. A Survey on Automatic Speech Recognition. IEICE TRANSACTIONS on Information and Systems, E85-D, 3 2002), 465-486.

[34] National Research Council. Educating Children with Autism. Division of Behavioral and Social Sciences and Education, Washington, DC: National Academy Press, 2001.

[35] Parés, N., Carreras, A., Durany, J., Ferrer, J., Freixa, P., Gómez, D., Kruglanski, O., Parés, R., Ribas, J. I., Soler, M. and Sanjurjo, A. Promotion of creative activity in children with severe autism through visuals in an interactive multisensory environment. In Proceedings of 2005 conference on Interaction design and children (Boulder, Colorado, 2005). ACM Press, New York, NY, 2005.

[36] Paul, R. Recommendation for Benchmarks - Autism Speaks Luncheon at SRCLD. L. DeThorne. 2008.

[37] Paul, R., Chawarska, K., Fowler, C., Cicchetti, D. and Volkmar, F. Listen My Children and You Shall Hear": Auditory preferences in toddlers with autism spectrum disorders. Journal of Speech, Language, and Hearing Research, 50. 2007), 1350-1364.

[38] Russo, N., Larson, C. and Kraus, N. Audio–vocal system regulation in children with autism spectrum disorders. Experimental Brain Research, Vol. 188, 1 (June 2008), 111-124

[39] Schneiderman, B. The eyes have it: a task by data type taxonomy for informationvisualization. In Proceedings of the IEEE Symposium on Visual Languages (Boulder, CO, 1996), 1996.

[40] Shuster, L. I. and Ruscello, D. M. Evoking [r] Using Visual Feedback. American Journal of Speech-Language Pathology, 1, May 1992), 29-34.

[41] Strik, H. and Cucchiarini, C. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication, 29, 2-4 1999), 225-246.

[42] Tartaro, A. Storytelling with a virtual peer as an intervention for children with autism. SIGACCESS Access. Computing(vol. 84 2006), 42-44.

[43] Tartaro, A. and Cassell, J. Playing with Virtual Peers: Bootstrapping Contingent Discourse in Children with Autism. In Proceedings of the Proceedings of International Conference of the Learning Sciences (Utrecht, Netherlands, June 24-28, 2008). ACM Press, 2008.

[44] Woods, J. J. and Wetherby, A. M. Early Identification of and Intervention for Infants and Toddlers Who Are at Risk for Autism Spectrum Disorder. Language, Speech, and Hearing Services in Schools, 34, July 2003 (July 2003), 180-193.