

© 2008 Joshua Hailpern

THE SPOKEN IMPACT PROJECT: USING AUDIO & VISUAL FEEDBACK
TO IMPACT VOCALIZATION IN NON-VERBAL CHILDREN
WITH AUTISTIC SPECTRUM DISORDER

BY

JOSHUA HAILPERN

B.S., Carnegie Mellon University, 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2008

Urbana, Illinois

Adviser:

Professor Karrie G. Karahalios

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 research, called the *Spoken Impact Project (SIP)*, we explore the effects of audio and visual feedback on vocalization in low-functioning children with ASD. By presenting a child with a new interpretation of their vocalizations (through audio and visual feedback), we aim to provide them with additional means of understanding and exploring their own voice. The SIP research spans over 12 months, including the creation of multiple software packages and detailed analysis of more than 20 hours of experimental video. This thesis details the four major components of this research project; 1) theory for visuals as feedback; 2) Supporting Video Annotation; 3) Creation of a Coding Guideline for Working with pre-verbal children and computers, and; 4) exploring SIP in an Experimental Context. In this work, we demonstrate the potential of computer generated audio and visual feedback to shape vocalizations of children with ASD.

ACKNOWLEDGEMENTS

It is because of the constant support, guidance and compassion of my advisor, Karrie Karahalios, that this research could even be conducted. I would like to thank all of my collaborators Jim Halle, Laura DeThorne, Mary-Kelsey Coletto, Christine Jakob, Ashley Sharer, Joey Hagedorn, and all of the students and faculty in the UIUC HCI research groups. Your insights, advice and constant encouragement have been critical. I also need to thank my parents, who have taught me both the beauty of life, and impact of technology. It is because of you that I see the potential of technology to bring about change in the lives of those who need it most. For this, and so much more, I will forever be grateful. Lastly, I must thank Marina. You keep me honest, critical, thoughtful, but most of all happy. Don't ever stop.

TABLE OF CONTENTS

Chapter 1: Introduction.....	1
1.1. Scope and Motivation.....	2
Chapter 2: Literature Review.....	5
2.1. Computer Visualizations.....	5
2.2. Autistic Spectrum Disorder	6
2.3. Communication Treatment.....	7
2.4. HCI & ASD Research.....	8
Chapter 3: Spoken Impact Project (SIP).....	10
3.1. Spoken Impact Project Software (SIPS)	10
3.1.1. Forms of Visual Feedback.....	11
3.1.2. Forms of Audio Feedback.....	13
3.1.3. Implementation	14
3.2. Research Questions.....	14
3.3. Within-Subject Experimental Design	15
3.3.1. Room Setup.....	17
3.3.2. Independent Variables.....	18
Chapter 4: Tools for video annotation	19
4.1. Related Video Annotation Work	21
4.1.1. Video Coding in Practice	21
4.1.2. Video Coding Tools.....	21
4.2. Interviews and Collaboration	23

4.3. VCode and VData.....	26
4.3.1. VCode	27
4.3.2. VCode Administration Window	31
4.3.3. VData	31
4.3.4. Implementation	35
4.4. Meeting the Requirements.....	35
4.5. Initial Reaction	36
4.6. Other Applications.....	37
4.7. Summary and Future Improvements.....	38
Chapter 5: A ³ Coding Guidelines.....	40
5.1. Related Work	41
5.2. A ³ : Refinement Methodology	43
5.2.1. Process of Refinement	43
5.3. A ³ Variable Description	44
5.3.1. Motor Behavior Variables	45
5.3.2. Verbal Variables	47
5.3.3. Other Metrics	49
5.4. Resulting Agreement.....	50
5.4.1. General Agreement & Accuracy	52
5.5. Discussion	53
5.5.1. Difficulties in Agreement.....	53
5.5.2. Other Feedback & Observations.....	55
5.6. Forms of Use.....	56
5.6.1. Research External to ASD.....	57

5.6.2. Use of the A ³ Guidelines	57
5.7. Summary.....	57
Chapter 6: Data Analysis	59
6.1. Dependent and Independent Variables.....	59
6.2. Question Analysis.....	60
6.2.1. Q1 Analysis	60
6.2.2. Q2 Analysis	60
6.2.3. Q3 Analysis	61
6.2.4. Q3a Analysis	61
6.2.5. Q4 Analysis	62
6.2.6. Q5 Analysis	62
Chapter 7: Results & Discussion	63
7.1. Subject 1: Oliver.....	64
7.1.1. Oliver's Results.....	64
7.2. Subject 2: Frank	65
7.2.1. Frank's Results	65
7.3. Subject: Larry.....	67
7.3.1. Larry's Results.....	67
7.4. Subject: Diana	69
7.4.1. Diana's Results	69
7.5. Subject: Brian.....	71
7.5.1. Brian's Results	71
7.6. Discussion	72
7.6.1. Q1 Discussion	72

7.6.2. Q2 Discussion	73
7.6.3. Q3 & Q5 Discussion	73
7.6.4. Q3a & Q4 Discussion.....	73
7.6.5. Parental Response	74
Chapter 8: Follow up study.....	75
8.1. Wizard-of-Oz Follow Up Study.....	75
Chapter 9: Conclusion	77
9.1. Limitations	77
9.2. Future Work	78
9.3. Summary.....	79
References	81
Appendix A: A ³ Coder's Guide.....	89
Appendix B: Letter of Copyright Permission.....	90
Author's Biography	91

CHAPTER 1: INTRODUCTION¹

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” [64], 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 [25], highlight the need for effective methods to facilitate the development of communication, including speech.

This thesis 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, generated from 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).

¹ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

SIP covers 3 major pieces of research; VCode and VData tools for Video Annotation and Analysis, A³ Coding Guidelines, and the Experimental Testing of SIP's Visualization software. In order to motivate SIP, we first outline the background, theory and existing state of both HCI and Communication literature. We then break SIP down into four major sections; Software & Experimental Design, Tools for Video Annotation, Dependent Variables for Behavior Analysis, Within-Subject Experimentation, and Data Analysis. When appropriate, we provide additional theoretical background within each sub section.

Beyond the experimental results, the contributions of this work are the creation of design guidelines to facilitate video annotation in an experimental setting based on the needs of researchers, a set of dependent variables to enable behavior analysis of non-verbal human-computer interaction, a 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.

1.1. 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 [91]. ASD

treatment research in HCI has targeted higher functioning children with ASD [101] but has failed to address the needs of non-verbal/low-functioning children with ASD. Though the literature in the behavioral sciences has explored this demographic, existing practices use low-tech alternatives such as PECS [19], mirrors and echo chambers [68] or invasive procedures, such as electropalatography [23]. 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 discussion that high-functioning children with ASD should not be pressured to communicate vocally, this concern is not applicable to this vein of research. These children cannot communicate by any means (e.g., typing, signing or speaking). Teaching some form of communication is essential, though the method should vary according to individual preference and capabilities.

The remainder of this thesis is structured as follows.

Chapter 2: Literature Review, presents a brief overview of ASD, computer visualization, and communication treatment.

Chapter 3: Spoken Impact Project (SIP) discusses our experimental & software design as well as a motivation for conducting this research project.

Chapter 4: Tools for video annotation presents two novel tools created to help researchers annotate video of experimental sessions in order to gather data that can be quantitatively assessed.

Chapter 5: A³ Coding Guidelines describes the creation and reliability of a set of dependent variables for assessing the interaction between non-verbal subjects and computer based feedback systems.

Both chapters present tools that were researched, designed, and created for support of SIP while having implications for use in other research and multiple fields. These two chapters provide additional literature reviews as they pertain to aspect of video annotation, and behavior analysis.

Chapter 6: Data Analysis and Chapter 7: Results & Discussion present the findings of this thesis, and a discussion of their significance.

Chapter 8: Follow up study presents the findings of a Wizard of Oz study.

Chapter 9: Conclusion summarizes the findings of SIP.

CHAPTER 2: LITERATURE REVIEW²

We present here an overview of existing research in Communication Treatment and HCI literature on Autistic Spectrum Disorder. Additional literature analysis is presented at the beginning of Chapter 4: Tools for video annotation, and Chapter 5: A³ Coding Guidelines, providing additional related work pertaining specifically to the topics covered in those chapters.

2.1. Computer Visualizations

Computer based visualization systems have been providing users with new, and faster ways to understand large quantities of varied and complex data [40]. See Figure 1 for examples of existing Visualizations. Work on Awareness Displays [67, 81] has shown the benefit of abstract representations, and their ability to provide continuous streams of data to users about their world, interests, and needs. Visualizations have also been used to impact teaching of complex concepts [61] by presenting visual and interactive representations.

Yet a major thrust of visualization research has focused on providing feedback to impact communication [14, 15, 46-48, 62, 103] and illustrate emotion [4]. These systems give the users a new understanding of their vocal interaction with others (from rate of speech to dominance in a conversation). These visualizations have the ability to alter people's behavior in

² Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

real time. This research has shown that these visualizations directly impact the communication of said individuals. Thus, the potential impact of visualization systems is far reaching, for social, personal, and educational purposes. We hope these ideas generalize to

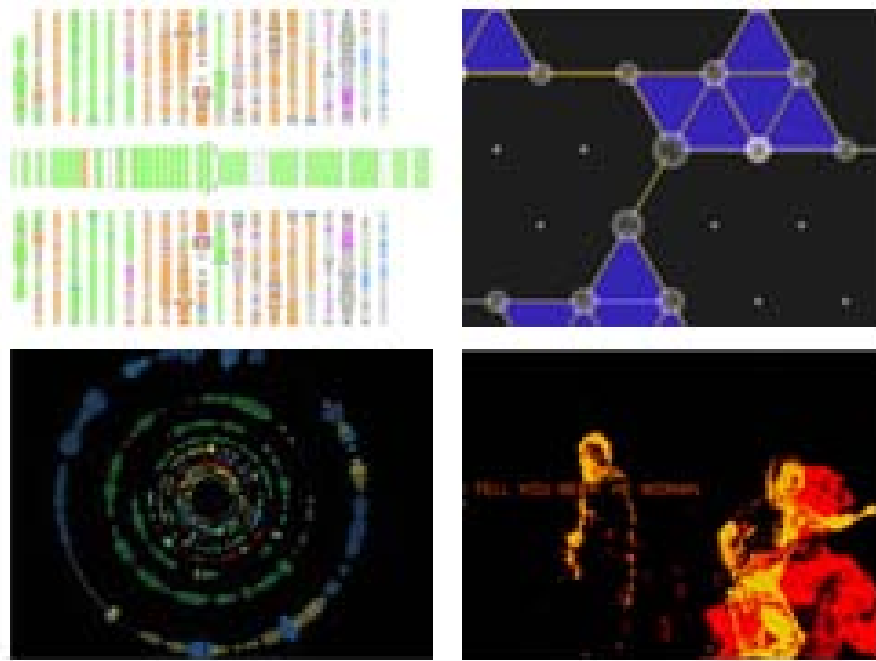


Figure 1. Examples of existing visualizations. Clockwise from top left: 1) Conversation Votes [14] 2) Isochords [15] 3) Conversation Clock [13] 4) Telemurals [45]

other communities and those with communication impairments.

2.2. Autistic Spectrum Disorder

Kanner's 1943 description [44] 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”[7]. While some children have limited impairment, those with a greater difficulty with social and communicative skills are considered low functioning.

2.3. Communication Treatment

Since the 1960s, Ivar Lovaas’ pioneering approach of “applied behavior analysis” has been used 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 [64]. Over time, rewards are slowly faded or removed resulting in more naturalistic behavior.

While the merits of this treatment have been documented for 30 years, this form of therapy has high financial and labor-intensive costs. Furthermore, frequent sessions requiring sustained attention and intense human-to-human contact can be anxiety producing [12]. This



Figure 2. Examples of existing communication treatment solutions. Clockwise from top left: 1) PECS [19] 2) Go Talk [6] 3) Speech Viewer III [42] 4) Visi-Pitch [49]

anxiety along with the detached/alone feeling of many children with ASD [12, 44] causes difficulty for practitioners and subjects. Further challenges also concern generalization of these skills. Other forms of communication treatment [35, 55, 75, 106] have been used to help develop social and communicative skills in children with ASD. Figure 2 illustrates four commonly used devices to assist in the development of communication skills.

2.4. HCI & ASD Research

Since the 1990s, the HCI community has examined how computers can aid in diagnosis of ASD [38, 52, 53]. In addition HCI has studies audio perception [91] and teaching human-to-human interaction to high-functioning children with ASD [51, 59, 72, 102]. Elements of play have also been studied that demonstrate that technology/computers can reduce the apprehension caused by human-to-human interaction [59, 70, 78]. Other HCI research [15, 42] and technology-based behavioral science research [3, 11, 97] outside of the ASD community has illustrated the use of computer solutions in the context of speech and communication therapy. Two of these solutions are illustrated in Figure 2. These computer-based solutions tend to approach visualization through games, and controlling characters in environments. While these game-like solutions can be useful, they abstract away many of the vocal properties resulting in a lack of understanding what or how their voice may impact the visualization. Further, they require a cognition level that can understand a game, a goal, and discovering how to achieve that objective.

Speech recognition is a commonly used technique for computationally capturing speech for the purposes of archival and analysis. Due to the current limitations of speech recognition

software [74, 98], the forms of speech detection are limited, especially for individuals with poor diction. Hence, technology must be designed to aid and supplement practitioners and researchers rather than replace them.

With this work, we explore methods and technology that can facilitate the speech and vocalization education process for children with communication skill deficits. Specifically we intend to use contingent visual and auditory feedback to (a) motivate and reward vocalization and (b) provide information about the acoustic properties of vocalizations. The following chapter details our approach for creating computer based feedback systems to encourage vocalization in non-verbal children with ASD.

CHAPTER 3: SPOKEN IMPACT PROJECT (SIP)³

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. This thesis presents the *Spoken Impact Project* (SIP) which examines 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 shape vocalizations of children with ASD.

3.1. Spoken Impact Project Software (SIPS)

During three months (Summer 2007), researchers designed the Spoken Impact Project Software (SIPS) package in Java using the Processing Library [32]. 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. We explored visual, auditory and mixed (both visual and auditory) feedback due to cross-modal interference

³ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

[20], commonly associated with ASD. Unlike existing software based communication treatments, discussed in the previous chapter, SIPS provides feedback and abstract representation of voice, rather than situating the visualization in a game with a concrete goal or objective.

We describe here the creation of SIPS, the metaphors of feedback, specific forms of feedback within each metaphor, and the implementation of the software package.

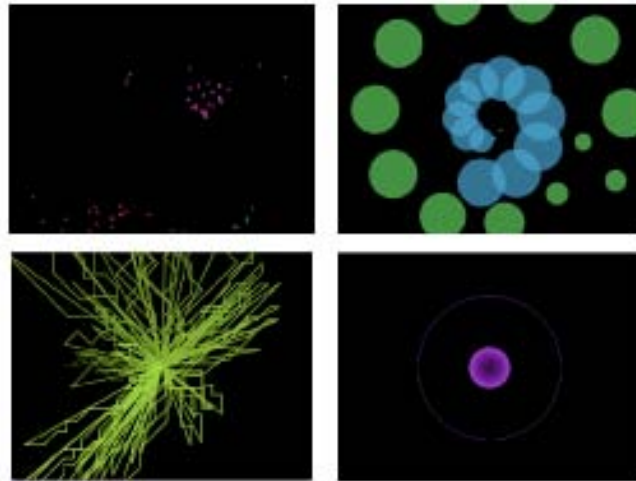


Figure 3. Examples of visualizations used in SIPS.

3.1.1. *Forms of Visual Feedback*

SIPS visual feedback (Figure 3) 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; Falling, Spinning, Flashing and Stationary. Among the four categories, approximately 12 unique motion/pattern

combinations were created; most can function with any type of object (circle, found image, etc).

3.1.1.1. Falling Visual Feedback

The falling metaphor simulates a gravitational pull on objects. This includes particle effects like water from a shower head or fireworks (Figure 3, top left). Consider a series of spheres falling, like rain, from the top of the screen to the bottom. If each sphere's diameter represented the volume of a sound, a subject could gain a temporal understanding of what sound they have made, and how their current sound differs. The falling metaphor was selected to leverage stimuli that garner interest from children with ASD [8, 34, 73].

3.1.1.2. Spinning Visual Feedback

The spinning metaphor moves objects in a circular or spiral pattern, outward, from the center of the screen (Figure 3, top right). Consider a series of circles (with a diameter based on the average volume in the past 2 seconds), which travel outward on a spiral path. The current sound remains in the center, and the subject can examine how their current sounds directly impact the circles which travel in a spiral. The spinning metaphor was selected to leverage stimuli that garner interest from children with ASD [8, 34, 73].

3.1.1.3. Flashing Visual Feedback

The flashing metaphor focuses on objects appearing and disappearing, giving a flashing or flickering effect (Figure 3, bottom left). Consider color circles appear in random locations on the screen, and the size of the color relates to the volume of the screen. This feedback would give an immediate sense of how loud a vocalization was, by the amount of color quickly

appearing on the monitor. Flashing feedback was investigated due to its high energy, which often appeals to neurologically typical children.

3.1.1.4. Stationary Visual Feedback

The stationary metaphor (Figure 3, bottom right) contained a singular object, at the center of the screen, and changing in one respect based on the sound of a subject. Stationary objects were explored to focus on change in an object (size, color, etc.) rather than object motion.

3.1.2. Forms of Audio Feedback

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

3.1.2.1. 1-to-1 Audio Feedback

1-to-1 feedback is 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). Though there was a slight delay between source sound and feedback, but both input and output occur simultaneously. By modifying a subject's sound, and returning it back to them, the impact of different vocalizations can produce a variety of new sounds, which can encourage vocalization.

3.1.2.2. Reward Based Audio Feedback

Computer sounds were 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). By treating audio as a form of reward, subjects were encouraged to produce sounds of a variety of length, in order to create audio that they enjoyed hearing.

3.1.3. Implementation

SIPS was built using Java 1.4 and the Processing Visualization [32] toolkit. During testing and experimentation, SIPS was run on an iMac computer, with the Phonex Audio SOLO microphone (noise canceling).

3.2. Research Questions

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

Q1: Will at least one form of real time computer-generated feedback positively impact the frequency 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 low-tech alternatives (e.g., image cards [19], mirrors [68]) and other related work. The remaining research questions examine modes of feedback, and their implications on frequency of spontaneous speech-like vocalization. Q2-Q5 are derived from research into cognitive profiles of children with ASD [60, 80] concluding that individuals with ASD prefer visual feedback [8, 34, 71, 73]. The responses to Q2-Q5 will directly impact future systems and the extent to which individualization is needed.

Q2: Will all forms of feedback positively impact the frequency of spontaneous speech-like vocalizations?

Q3: Will subjects increase the frequency of their spontaneous speech-like vocalizations in all conditions with visual only feedback, audio only feedback and/or mixed feedback?

Q3a: If there is a modality that approaching or is significant (Q3), is there a specific form of that feedback in that modality that positively impacts frequency of spontaneous speech-like vocalizations?

The quantitatively driven investigation of Q3 may hide the impact of a specific form of feedback. If that one form of feedback fails to significantly adjust the results in Q3, it will never be analyzed in Q3a. Therefore;

Q4: By testing feedback conditions that were qualitatively favored by subjects (assessed during experiment and via video), will we uncover forms of feedback that positively impact the frequency of spontaneous speech-like vocalizations?

Q5: Is there a modality of feedback whose variations indicate (Q3, Q3a, and Q4) the child's frequency of spontaneous speech-like vocalization are positively impacted.

3.3. Within-Subject Experimental Design

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 for comparison. 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 [50], an adaptation of the alternating treatments design [10], 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.

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 different forms of feedback (permutations of audio and visual). Each trial began with an antecedent demonstration by the researcher (e.g., saying "boo" and pointing to screen). The subject then could engage the system in whatever manner they chose.

Feedback permutations were selected based on qualitative vocalization frequency. 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.



Figure 4. 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

3.3.1. Room Setup

Due to the varying personalities and ability to attend in a chair for an extended time, a variety of room configurations were employed. The room configuration was selected based on each child's preference and ability to sit in a chair, assessed during the orientation sessions. Figure 4 illustrates the room configurations.

3.3.2. *Independent Variables*

Our within-subject experiment analyzed the dependent variable Spontaneous Speech-Like Vocalization (SSLV). A more detailed explanation of dependent variable selection is conducted in Chapter Chapter 5: A³ Coding Guidelines. The independent variables were the various permutations of visual and auditory feedback. This facilitated contrast between the mode of feedback (visual, auditory, and mixed) as well as the different types of feedback (12 visual and 5 auditory forms).

This chapter describes the methodology, software technology and execution of SIP. The following chapter presents a set of guidelines to design tools for video annotation and an instantiation of those guidelines as a software suite called VCode and VData. Through the use of VCode and VData, researchers are able to collect large quantities of data to analyze the quantitative impact of SIP.

CHAPTER 4: TOOLS FOR VIDEO ANNOTATION⁴



Figure 5. VCode and VData Suite of Applications for Video annotation

Human behavior does not naturally lend itself to being quantifiable. Yet time and again, researchers in disciplines ranging from psychology to ethnography to computer science, are forced to analyze as if it was quantified. Those in human centered domains can now rely on video annotation to provide them with measures on which to draw conclusions. Unlike transcription, which is akin to what a court stenographer does, annotation is the marking of movements, sounds, and other such events (with or without additional metadata such as

⁴ Some research, content, and text from this section is reproduced with permission from the Association for Computing Machinery [Appendix B].

Hagedorn, J., Hailpern, J. and Karahalios, K. G. VCode and VData: Illustrating a New Framework for Supporting the Video Annotation Workflow, *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2008)* pp. 317-321. <http://doi.acm.org/10.1145/1385569.1385622>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. AVI '08 May 28-30, 2008, Naples, Italy Copyright 2008 ACM 1-978-60558-141-5 ...\$5.00.

rankings). The emergence of technology as a tool to aid in video annotation has raised the possibility of increasing reliability, repeatability, and workflow optimizations [21].

Three notable limitations of existing video annotation tools are lack of support for the annotation workflow, poor representation of data on a timeline, and poor interaction techniques with video, data, and annotations. This chapter details a set of requirements to guide the design of video annotation tools. Our model is the direct result of an analysis of existing tools, current practices by researchers, and workflow difficulties experienced by real-world video coders. By understanding what data researchers are looking to gather, and the shortcomings of existing techniques and technology utilized by coders, we believe that we have created a framework for video annotation that can reach across disciplines. Our model is demonstrated through the design and construction of our new system VCode and VData (Figure 5); two fully functional, open-source tools which bridge the video annotation workflow.

The primary contribution of this chapter is the set of design requirements for facilitating a system conducive to video annotation. Specifically, we demonstrate how a system could be designed and built to meet these requirements through a set of carefully designed interfaces and graphical representations of data.

4.1. Related Video Annotation Work

4.1.1. Video Coding in Practice

The analysis of human behavior is a study that dates back hundreds of years. This has ranged from anthropological ethnographies [27] to psychological evaluations. As technology has developed, the use of video and creation of annotation techniques have aided researchers by providing a referable document that can be used as evidence to back up claims and observations made [58, 89, 90]. These techniques involve detailed event logging on paper, specifying features such as durations, ratings/levels, and time-stamps [57]. To ensure a reliable set of data from annotation, researchers perform agreement calculations between coders [16]. This agreement is utilized throughout the data gathering process (by testing some small percentage of data segments to ensure consistency throughout), but also during training of coders (to decide when they fully understand what events they are looking for). There are many techniques for calculating agreement including Cohens Kappa [29, 50], Cochran's Q-test, and Point-By-Point Agreement. Regardless, the management of data with traditional means is considered "cumbersome" [89].

4.1.2. Video Coding Tools

Digital annotation tools have demonstrated significant benefits from simple copy/paste and undo to increased quality of coding by the facilitation of multiple passes on video and graphical representations [21, 84]. A timeline is commonly utilized in these tools and is familiar without extensive training [30]. Existing research also indicates that presenting coders with secondary or sensor data on a timeline helps them outperform coders without sensor data [21]. Increased

accuracy, quality, and speed not only enhance the data collected, but also allow for more annotation to be conducted in the same amount of time. In addition to the computational benefits of digital annotation tools, they also provide a controllable mechanism for deferent forms of reliable video playback. [84].

One critical limitation of existing tools is poor representation of data on a timeline and utilization of screen real estate. For example, the VACA solution, while utilizing minimal screen real-estate by condensing all annotations to one large easy to read track, presents a problem with overlapping and simultaneous events [21]. The VisSTA solution takes the contrary approach by showing many small vertically tiled tracks. Though this allows for a good comparative view, reading individual annotations & holistic interpretation is difficult due to scrolling [84]. These and other existing solutions have not successfully dealt with this problem [2, 30, 43, 54, 76, 85, 94, 99].

Another limitation of current annotation tools is poor interaction techniques with video and data. Though robust functionality is provided for playback, controls can be cumbersome & overly complex, (e.g. [84]). Too many windows resulting in an over-saturation of information, imprecise video interaction & annotation or rigid, inaccessible marking interfaces (e.g. [54, 82, 84, 85, 99]). Each of these are common stumbling blocks which could result in unreliable data.

One last limitation is lack of support for the full annotation workflow that follows researcher from experimentation to data analysis. The larger the degree of support, the smaller the chance of error, and the more efficient the data gathering, collection, and analysis process

becomes. This workflow was created through discussions with coders, researchers and an examination of the existing literature. The following are the 6 steps of the video annotation workflow:

- 1) collect video
- 2) create segments to code
- 3) train coders/demonstrate reliability
- 4) gather data
- 5) perform regular checks on reliability & discuss discrepancies
- 6) perform data analysis.

Many tools support small portions of this workflow (i.e. simply facilitating segmentation, annotation, or reliability [21, 94, 99]), but with each break in the process researchers can become delayed. Without export/import data reentry is required. Technology is situated to optimize this process. Researchers have also explored dialogue transcription [43, 93, 99], tagging [2, 99], scene based automatic annotation [26, 82], automatic event logging [9], and object of focus identification [17]. This chapter contrasts these other foci by demonstrating techniques for supporting human based annotation of events that occur in video.

4.2. Interviews and Collaboration

To gain a deeper understanding of methods, analysis processes, bottlenecks, and types of data needed for effective video annotation software, we maintained an active dialogue with researchers (in Special Education, Speech and Hearing Sciences, and Computer Science) who use video annotation, conducted informal 40 minute interviews with two experienced video coders, and refined functionality through dialog with current users of VCode and VData.

Existing tools for video annotation may address a subset of the below described requirements, however, our system more fully satisfies all of them.

R1: *Facilitate Coding Workflow:* The coding workflow consists of; (1) establishing video clips and coding guidelines, (2) intense training of coders and checks for reliability, (3) annotation of videos, (4) weekly reliability checks on annotated videos, (5) repeat 3 and 4 ad infinitum, (6) analyze data in statistical packages. Tools targeting video annotation should attempt to optimize the transition between steps in this workflow.

R2: *Video, Annotations, and Coding Guidelines should be presented in a synchronized manner:* Interviewees described their coding process centering around analog video on a TV-VCR device, annotating in a Microsoft Excel file, and referencing lengthy code guidelines. Due to the visual separation between annotations, source material, and video, coders had great difficulty during reviews.

R3: *Capture Appropriate Data:* Researchers and existing literature indicate that there are different types of data that are collected through the annotation process: counting events/occurrences, determining duration of events, assigning levels, values, or ranking to events, performing phonetic transcription, and general commenting [54]. Effective interfaces must provide methods for capturing these conceptually different data types while preserving each of their unique nuances.

R4: *Additional data should be displayed to coders:* Effective annotation tools should allow researchers to provide additional data to coders to aid in their assessment of video; for example, a volume histogram of the current video, sensor/log data collected in tandem to the video capture, or annotations

made automatically or from another source. Displaying additional datapoints has shown to increase the accuracy of coded events [21]. Further, annotation software should facilitate the management of multiple video streams to get the most accurate "view" on the session, and thus produce the most accurate data [84].

R5: Allow multiple forms of playback: Researchers mentioned that continuous playback is not always the preferred method of analyzing a video. Often multiple modes of playback are utilized; continuous or standard playback, continuous interval playback (play for N seconds, then stop), and skip interval playback (jump N seconds, then stop). This allows the video to be divided in to smaller segments for annotation of events that are more difficult to pinpoint (i.e. when a smile starts or ends) [84]. Though conceptually simple, manipulations of video using a standard VCR was described as "annoying" and "a mess" due to hand eye coordination and repeatability issues.

R6: *Agreement calculations should be easy and manipulatable:* Regardless of agreement technique used, researcher expressed a frustration in attempts to calculate inter-observer reliability. Specifically, existing solutions were limited to importing data into a statistical software package for calculation or calculating them by hand. Video annotation tools should provide quick & easy reliability calculations for individual variables, as well as overall.

R7: *Provide functionality for visual, graphical and contextual review of annotations:* In interviews, coders lamented the process of ensuring reliability on a weekly basis; as it consisted of searching through printouts of a spreadsheet for discrepancies. Specifically, by lacking context in this spreadsheet coders found it difficult to recognize what a given coding mark referred to due to the lack of synchronization with video. By providing a visual, graphical way to

review annotations (in the context of the video) coders would be better able to justify the decisions, determine the correct solution, and save time identifying the errors.

4.3. VCode and VData

VCode and VData are a suite of applications which create a set of effective interfaces for the coding workflow following the above design requirements. Our system has three main components: VCode (annotation), VCode Admin Window (configuration) and VData (examination of data, coder agreement and training). The interaction with VCode and VData is demonstrated in Figure 6 - Figure 11 in which two coders are marking a video of a child in an experiment, and checking the agreement between their annotations. The reader should note our solution is only one possible implementation of the design requirements, and that these requirements could be applied to improving existing video annotation software.



Figure 6. Two Coders and Researcher reviewing a coding session

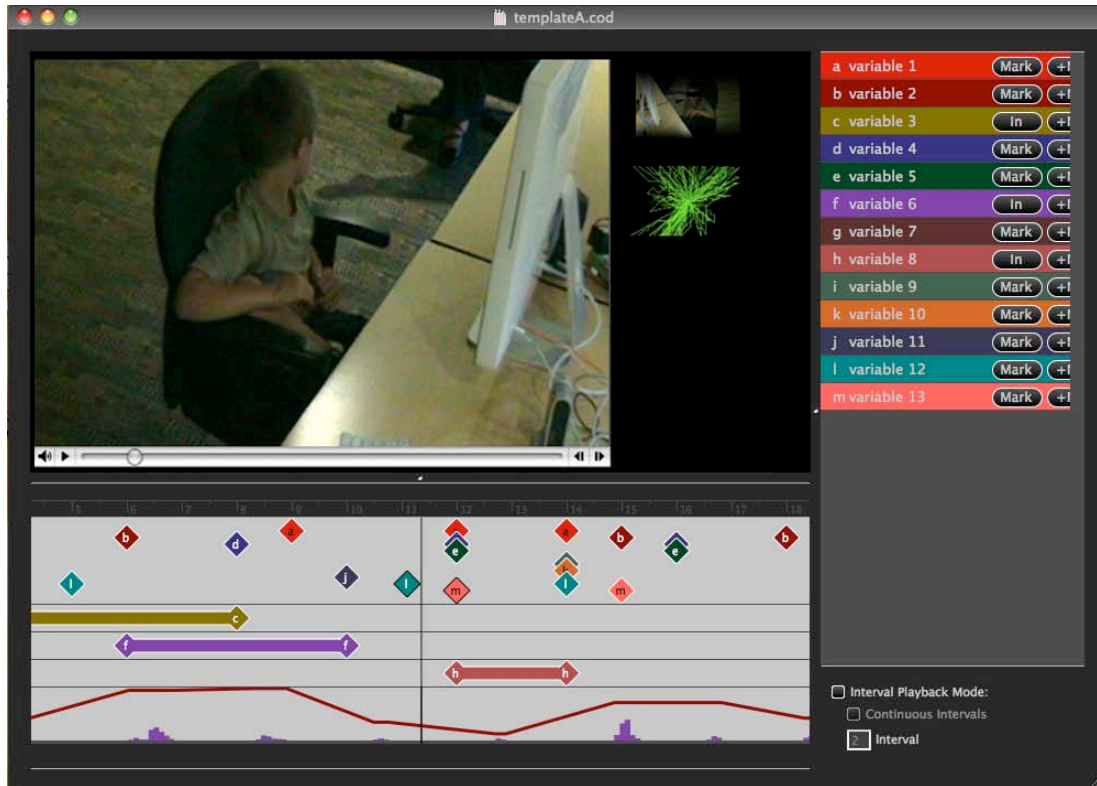


Figure 7. The VCode application graphically represents the behaviors being coded as marks on a timeline. It is easy to see correlation between the marks on the timeline and sensor data displayed below.

4.3.1. VCode

The VCode application (Figure 7) is designed to provide researchers with an effective way to obtain reliable data from an observational research video. By allowing researchers to present multiple video streams in addition to other sensor data (e.g. log data, annotations from other software, or signals recorded by a computer/monitoring device) the coder can make the best annotation decision possible.

Video: To facilitate multiple video streams VCode presents one main video at full size, and a dock with other streams playing in real time. When a docked stream is clicked on, it

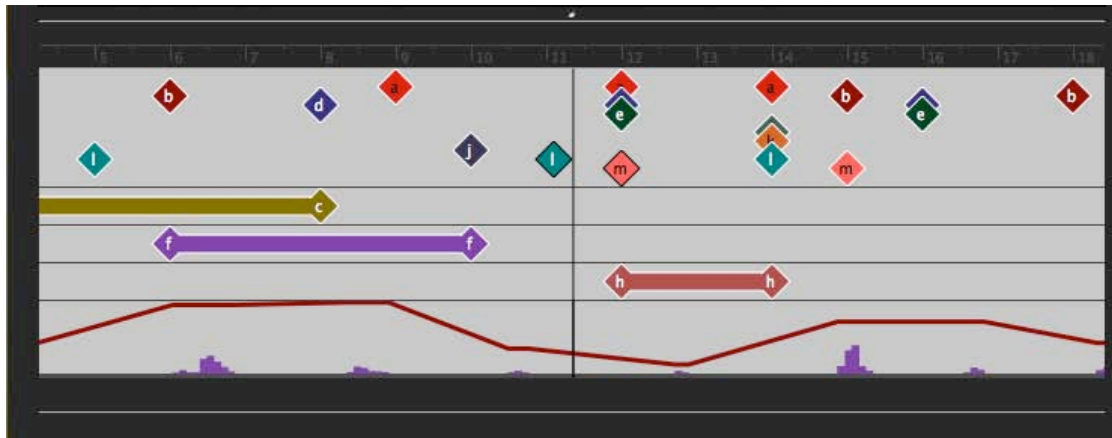


Figure 8. The VCode timeline. Each event is marked as a diamond, with the corresponding hot-key being represented on the mark for easy identification. A time stamp runs along the top of the timeline, with a play head represented vertically.

repositions itself into the main video window, while the video which was the previous focus, scales down to the dock, thus equating visual importance with relative size and visual weight.

Variable List: To provide coders with a persistent list/understanding of their task and objective, VCode presents them with a list of each variable they need to annotate. This list (presented along the right hand side of the VCode window) displays each variable's name, a unique color associated with that variable, a keyboard hot key for placing marks, as well as UI buttons to facilitate mark placement.

Events: When annotating a video, two different classes of coding events emerge: ranged and momentary. A ranged event is one that extends over a period of time (marking action start and duration). Momentary marks have no duration, and thus represent one specific moment in time. Each mark appears with it's associated color and it's keyboard hot-key letter (see Figure

8). These color and hot key based identifiers make it easy for coders to quickly assess which events represent which variable.

Comments can be attached to any mark, allowing additional observations, levels/ranking, or phonetic transcription (through onscreen phonetic keyboard). Any mark with a comment has an inverted outline to signify that it has a comment attached. Figure 7 and Figure 8 show a ranged event representing the length of time which a child is making a sound, with additional momentary marks at the start noting other features of the child's state of being).

Timeline: The timeline is the heart of VCode. It is modeled after the moving timeline one might find in a video editing application (e.g. iMovie, Final Cut Pro, etc.). Events, graphically represented by diamonds, appear in a spatial linear fashion to sync with the video. Once an event has been placed on the timeline, dragging, clicking, and double-clicking can graphically manipulate the mark. The standard solution for dealing with large numbers of tracks or variables is to provide a vertical scroll bar or overlay tracks. Rather than limiting the amount of information on screen by scrolling, tracks representing momentary events are "stacked," such that they vertically overlap. This optimizes usage of the screen while still providing enough area for track isolation and selection, even under dense data conditions.

Ranged event tracks are unable to benefit from this stacking optimization because of the more complicated interaction for manipulation and thus are vertically tiled. Researchers can present video volume, sensor data, software log data (from Eclipse or Photoshop for example), and even other annotations to the coders. This additional information is presented graphically to

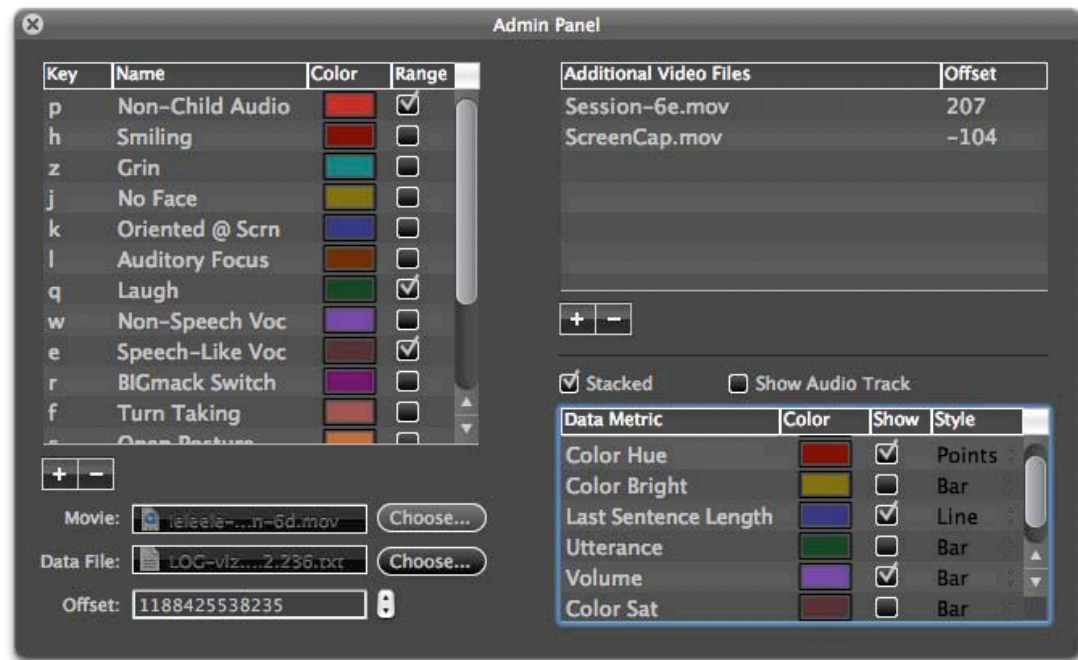


Figure 9. The code is specified in the Administration Window along with the different video angles, screen capture, and log data.

the users by bar, line, or scatter plot. This secondary data can allow coders to annotate data captured by other sources than the video streams, as well as provide additional context to their code. For example, should a coder be instructed to mark when a certain noise occurs, he can line the mark up with an audio peek, rather than estimate it and be concerned with reaction time.

Interaction: Annotations can be inserted into the timeline via UI buttons or keyboard hot keys. In order to increase association between hot-key and timeline, timeline marks have the letter of their hot-key on them. To optimize the typically complex transport controls we isolated the key activities that coders need execute and provided controls limited to play/pause

buttons, coarse and fine grained playhead positioning, and step controls. The three modes of playback outlined in R5 are available.

4.3.2. VCode Administration Window

To ensure consistent configuration between coders and sessions, all administrative features are consolidated in a single window. The expected workflow is such that a researcher would setup a single coding document with all the variables to be used on all the videos. This template would then be duplicated (with media and log files inserted for each trial). The main task the Administration Window (Figure 9) is to facilitate is the creation of tracks, used to code data. Researchers can add, remove, and reorder tracks that appear in a list format. The name, color and hot key of each track can be set through this list presentation. Tracks can be enabled as ranged events through a check box in this interface. The Administration Window is also where a researcher specifies videos and data file to be coded, as well as secondary data for contextual annotation. These elements are specified and synchronized through a drag and drop interface, all of which is hidden from the coder to prevent configuration corruption.

4.3.3. VData

Critical aspects of the video coding workflow (training, reliability, and accuracy) revolve around demonstrating agreement between coders. VData (Figure 10) is a separate executable application specifically targeted to aid researchers in training and agreement analysis of coded data produced in VCode.

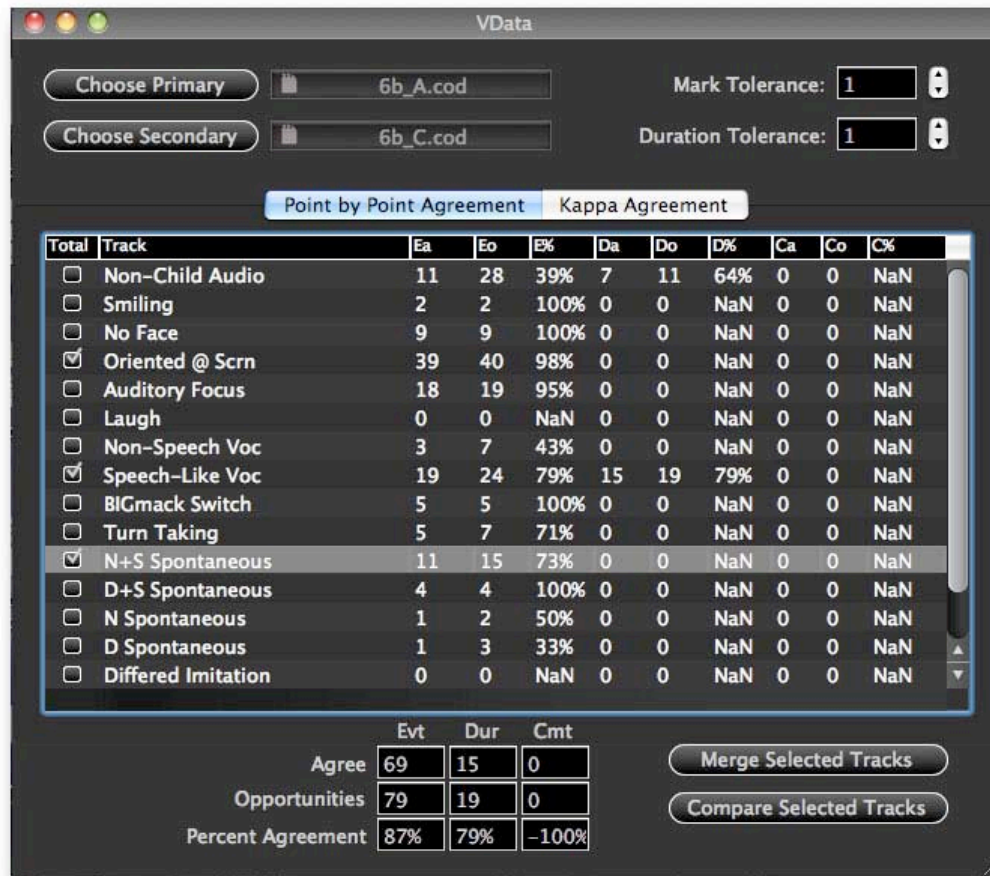


Figure 10. Later, analysis is performed on independent codings of the same video.

Multi Coder Analysis: By loading two VCode files into VData, tracks are automatically loaded into the main data table that presents all the raw data (opportunities, agreements) and percentage agreement for point-by-point agreements calculation. For each event (momentary or ranged) an opportunity is said to occur when the primary coder makes a mark. If the secondary coder also makes a mark within a specified short interval, the marks are said to agree. A percentage is calculated from $\frac{\text{agreements}}{\text{opportunities}}$ for easy interpretation. A tolerance variable is also present to (1) accommodate for variability in the mark placement by the coders, and (2) recognition that there is no quantization of marks beyond the granularity of the

millisecond timescale, a property of the system. VData also provides agreement for ranged events and annotations in a similar fashion.

In addition to point-by-point agreement calculations, VData also supports Kappa Calculations by switching tabs. Much like point-by-point agreement, all of the raw data needed to perform a Kappa calculation is presented in the application. Kappa calculations is based upon the notion that within a given time span (the video duration), there are a discrete segments called opportunities. By opportunity we mean, that in a certain period, an observation can be made (aka something DID or DID NOT occur). When calculated, Kappa takes into account the probability that two coders made a mark at the same location by chance. The presence of a annotation in VCode, means an event occurred. The absence of a mark means it did NOT occur. Kappa examines the number of agreements between two coders for events occurring and not occurring. As a result, this form of agreement calculation only applies to those tracks that were annotated using interval playback mode.

It is not uncommon for multiple tracks or variables to by measuring slight variations on a theme (e.g. smiling vs. large smile vs. grin), thus VData implements a track-merging feature which allows opportunities on two distinct tracks to be treated indistinguishably. The resulting hybrid track can be used to see all the same raw data in addition to the percentage agreement. This new track can be treated as any of the original tracks from the VCode files.

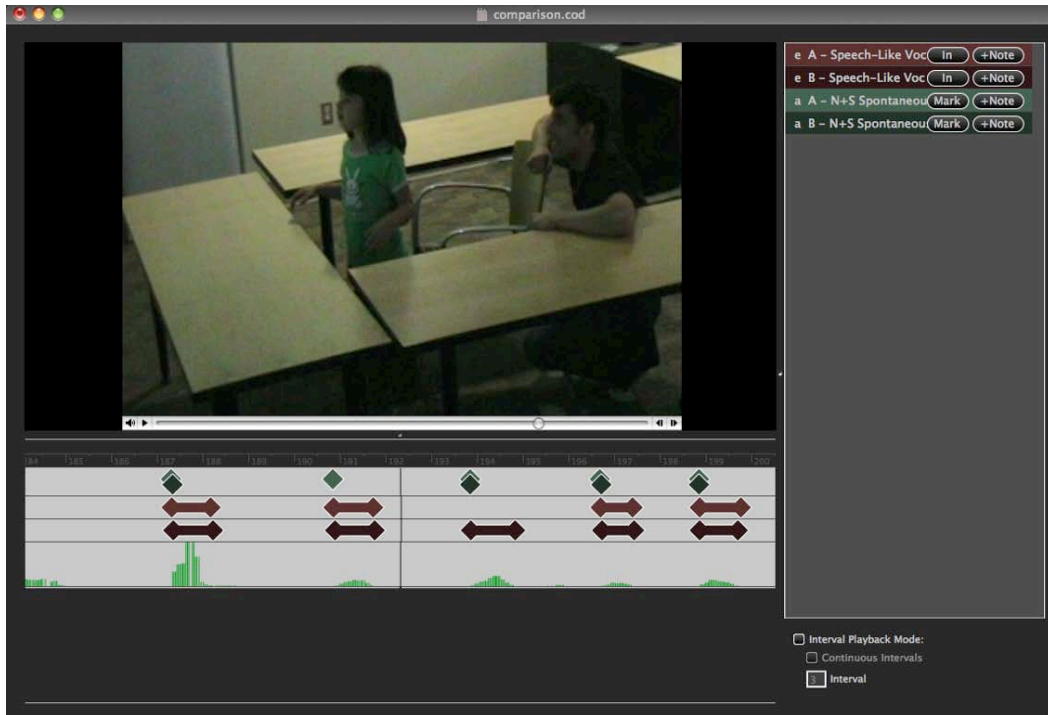


Figure 11. A track with low agreement can be reconciled by viewing the results of two coders side-by-side in VCode, thanks to the capabilities of the VData analysis tool.

For a holistic view, researchers can select tracks to be added into a total agreement calculation. In other words, if analysis determines that a single track is not reliable or it is determined that a given track will not be used in the future, it can be easily excluded from the total agreement calculation.

Conflict Resolution & Exporting: We have optimized coder training and reliability analysis by providing a graphical mechanism to directly compare annotations of two coders. VData can create a VCode session containing specific tracks of two individual coders for side-by-side comparison (Figure 11). Both coder's marks appear side-by-side along side the source video, both with similar colors (e.g. light green, and dark green). The visual, side-by-side,

representation of the data makes it easy to recognize systematic errors in context and detect differences between two coders markings. This reduces the time necessary to locate discrepancies and discuss the reasons why they might have occurred. It is necessary to keep records of these agreement analyses performed with VData by text export. Maintaining export at each stage of the process provides additional transparency and maintains traceability of results that come out of the system.

4.3.4. Implementation

VCode and VData were implemented in Objective-C using the Cocoa Framework for Mac OS X 10.5. VCode supports all video formats and codecs supported by QuickTime to enable wide compatibility with available video files.

4.4. Meeting the Requirements

To ensure Video, Annotations, and Coding Guidelines are presented in a synchronized manner, VCode provides a unified interface containing the target video, a timeline with graphically represented annotations (ranged event, momentary event, or comment depending on data metaphor) (**R3**), additional tracks of signal data (to increase accuracy) (**R4**), and a list of coding guidelines (**R2**) which place marks and stand as a visual reminder. Three forms of video playback (continuous/standard, interval playback, skip interval playback) are available via check boxes on the main VCode window to allow easily switching between modes of playback (**R5**). VData provides a dynamic interface for real time calculation of multiple agreement values to facilitate easy and dynamic agreement calculations (**R6**). Through the transparent calculation process, researchers can see both the raw data, and the percentages side

by side for easy judgments about the reliability of data collected. Upon request a visual, graphical and contextual review of annotations for both agreement review and training is supported (R7).

Finally, the Coding Workflow (R1) is encouraged through VCode's template model in conjunction with the separate VCode Administration Window for easy set up and configuration. Training, data collection, and inter-coder agreement are enabled through a tight collaboration between annotation environment and agreement analysis. By consistently providing data export, researchers can be assured that any information annotated by coders can be easily extracted and exported into the statistical analysis tool of their choice.

4.5. Initial Reaction

To evaluate our system in a cursory fashion, we conducted an informal series of interviews with several coders that used our system during the course of an independent study. Analysis using VData showed inter-observer agreement was good and provided valuable coded data for the study. In general, comments from the coders were positive, especially when comparing the VCode system to non-computerized methods. One coder wrote: *"The software was easy to use in general, and cut down on coding time."* Several features of VCode stood out in their comments; color coding of tracks provided direct linkage between events on the timeline and the description panel, the correlation between files was clear to see during review, sensor data helped anticipate events and accurately code them. It was also noted that the sensor data provided reassurance that what they had noticed in the video was actually correct.

In addition to these positive marks we uncovered several shortcomings of the interface. The seemingly low-resolution bar-graph of volume data left coders unsure where precisely to make their mark. Because the elements of this graph are relatively wide, it appears especially coarse in comparison with the precision with which one may place a mark on the timeline. A spectrogram was suggested as an alternate visualization of the audio data that could help understand sound and video.

From a quantitative standpoint, the time required to annotate video for SIP drastically improved over the course of the experiment. At the beginning of their coding experience, coders took roughly 40 minutes per 1 minute of video footage (to annotate all variables listed in Chapter 5: A³ Coding Guidelines). By the end of the coding period, coder's time was reduced to about 20 min per 1 minute of footage. Coders attributed this improvement to the ease of use of the annotation system.

Overall, results from these interviews and raw speed improvements, are very encouraging and suggest a more formal study to determine if performance improves in the same way that coders stated that they felt as the tool lowered the amount of time necessary for coding.

4.6. Other Applications

The VCode framework has additional applications outside of the HCI and Behavioral Science research communities. In many clinical settings (e.g. speech pathology), practitioners record their sessions with a subject, and post-hoc analysis of the video to assess subject's progress. Many of the same features provided in VCode can facilitate analysis of clinical sessions,

providing therapists with a quick, reviewable and accurate tool for documenting progress and behavior. Further, the ability to export data allows clinicians to graph progress made, and be able to show the client (or guardian) the beneficial effects of treatment.

4.7. Summary and Future Improvements

Video annotation tools can be valuable to researchers by enhancing the annotation process through increased reliability, repeatability, and workflow optimizations. However, many existing solutions do not fully address all the needs of researchers and coders; effective representation of data on a timeline, efficient and robust interaction techniques with video and data, and support for the full video annotation workflow. Our research has provided many contributions in addressing these weak points.

We create a set of design requirements based on existing literature and annotation techniques, interviews with experienced coders, and discussions with researchers in multiple disciplines. Based on these investigations, we implemented a system, VCode and VData, that largely satisfies the requirements we outlined. These systems were then used in SIP, and coders were interviewed concurrent with and after using the software, and their reactions were solicited. Our model demonstrates how video annotation software, for many disciplines, can be enhanced to meet the needs of both researchers and coders.

As of the date of this thesis, 9356 downloads of the VCode software package from multiple universities and countries around the world. Of note, the VCode package has been used at Departments of Computer Science, Human Computer Interaction, Psychology, Cognitive

Science, Neuroscience, Education, Information Systems Engineering, Design Research, and Library & Information Science.

From the reaction of the coders, as well as our own assessment of VCode and VData, we have many directions of possible future work. One avenue is creating a database or networked system in order to facilitate remote access to content, and management of coding objects and assignments for individual coders. It is foreseeable that the system could be extended to a tool to prepare coding files; assist in dividing up raw footage, syncing data to video enmasse, and other automation hooks. This could leverage some of the other existing work in automatic video segmentation. Lastly, we hope to address some of the concerns of our coders, including creating a richer set of data visualizations.

With such a tool sets available to researchers, more complex sets of variables are available to research to code. The follow chapter details the development of a new coding system that leverages many of the features of VCode and VData. This new coding guide allows researchers to assess the impact of computer based feedback systems, on nonverbal subjects through behavioral analysis.

CHAPTER 5: A³ CODING GUIDELINES⁵

Work conducted in HCI to date has explored diagnosis [52], play [70, 78], audio perception [91], and interpersonal skills for high functioning children with ASD [102]. Although this work is greatly beneficial, the potential of technology to facilitate vocal development in lower-functioning children with ASD has received little attention. As a result, there is little work in the HCI domain that provides a model for how to quantitatively assess the impact of an intervention to encourage speech with low-functioning children with ASD using HCI. It is essential, that whenever assessing a novel design approach, to have tools and methodologies (in the case of this research, in particular, and for assistive technologies, in general) to document that design and those techniques, so that the value of the novel approach can be evaluated and compared to the state of the art.

We propose A³ (pronounced A-Cubed) or Annotation for ASD Analysis to quantitatively assess a set of dependent variables identified through the digital video annotation process.

Through the application of A³ in this research context, we demonstrate the inter-rater reliability of the annotations, as well as directions for its improvement. Because we are

⁵ Some research, content, and text from this section is reproduced with permission from the Association for Computing Machinery [Appendix B].

Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. A3: A Coding Guideline for HCI+Autism Research Using Video Annotation. *Proceedings of the 10th Intl SIGACCESS Conference on Computers and Accessibility (ASSETS 2008)*, pp. 11-18. <http://doi.acm.org/10.1145/1414471.1414476>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. ASSETS'08, October 13–15, 2008, Halifax, Nova Scotia, Canada. Copyright 2008 ACM 978-1-59593-976-0/08/10...\$5.00.

required to rely entirely on subject behavior, rather than on feedback provided by subjects (due to the nature of ASD), the creation of such an assessment tool as A³ is critical for evaluation of technology used by the ASD community. The contribution of this chapter is in the demonstration of a new coding system grounded in theory from multiple domains and demonstration of its reliability when applied in the context of SIP, an experimental study.

5.1. Related Work

Analysis of human behavior can be traced back hundreds of years [27] from anthropologists examining the behavior of societies to psychologists who explored man's individual and social behavior. More recently, those using technology (Human Factors, Computer Science, CSCW) have also looked at human behavior. As technology advanced, new techniques for notation of behavior have emerged. With the advent of video, these forms of analysis became linked to replay-able clips, allowing annotations on specific actions to be made. These links to re-watchable video allowed researchers to quickly refer back to the actual events, rather than rely exclusively on notes and memory [58, 89, 90]. As this technology has evolved, varying guidelines and dependent variables have emerged to help shape research in different communities.

Those in the behavioral sciences have spent decades analyzing video and constructing coding schemes to allow investigators to better understand the impact of traditional therapeutic interventions. Some of these investigators analyze aspects of speech an/or sound production [56, 77, 104, 106] , interaction (with and without physical presence) [12, 104], diagnosis via observation [63] and communication skill acquisition [83, 96]. Each of these domains of work

relies on coding guides targeting aspects of behavior under naturalistic or analogue conditions. Such studies, which rely heavily on reliable coding, have not focused their publications on the coding methodology.

There is an interesting parallel between the subjects in Infant Research and those diagnosed as low functioning ASD in that both populations are non-verbal. In many respects, they present similar levels of communicative skills. Although our work has a different purpose and is in a different context, it shares many of the same critical aspects of behavior analysis with Infant Research [39, 66, 95].

Computer Scientists, particularly in HCI, have developed a broad set of coding guidelines for a large array of tasks [100, 105]. However, few of these focus on the evaluation of subjects who are non-communicative. Even fewer address those with ASD. There exist guidelines that have dealt with higher functioning subjects [24, 102] and most, gathered data through subjective (qualitative) observation [51, 70, 78].

Although the literature in these disciplines is comprehensive and examples of coding guidelines are robust, there currently does not exist an established quantitative coding system that addresses low-functioning children with ASD and interventions using computer systems that provide auditory and/or visual feedback. This chapter addresses this gap by detailing the construction of A³ and the reliability of the variables in an experimental research setting.

5.2. A³: Refinement Methodology

SIP was designed to examine children's responses to computerized feedback, in terms of their engagement, attention, and vocal behaviors. Although the primary mode of data collection was through video annotation, the precise definition and selection of dependent variables evolved over the course of the study. Variables were initially identified by analyzing relevant target behaviors in related disciplines (Speech and Hearing Science, Special Education, Psychology and Computer Science), by observations of investigators present during sessions, and by an examination of a small, random, subset of video from the sessions. VCode and VData were utilized for both data collection, and agreement calculations.

5.2.1. *Process of Refinement*

Although the initial set of coding guidelines was created during the experimental process, the final set presented in Appendix A and Section 5.3 were created through an extensive iterative design cycle. Starting mid-October 2007, researchers met weekly for seven weeks with two coders. During these meetings, trials from sessions were randomly selected and annotated collectively to further refine the definition of each dependent variable.

Following this initial phase of refinement, the same two coders were asked to annotate video from the same five random trials, once a week. Video was reviewed by researchers and coders for agreement each week, and further refinement of coding definitions ensued. In mid-December 2007, two new coders were introduced, trained, and eventually replaced the original pair. Although this shift was originally made due to personnel availability, adding new "eyes and ears" to the annotation process ensured that any assumptions about variable definitions

made by the first pair of coders were revealed and explicitly noted in the guidelines through several weeks of training/guideline-refinement. All coders were students in the Speech and Hearing Science Department; the first pair were seniors, while the second pair were graduates of the undergraduate program (one was pursuing a masters in Speech and Hearing Science). All video coders had class experience in phonetic transcription and three of the four had worked as coders on relevant research projects.

The refinement period was concluded when an 85% agreement criterion was met across all variables on one trial. Though 80% is 'generally' considered to be an acceptable rule of thumb [50], we wanted to ensure that agreement was above 'standard'. Agreement for this research was determined by point-by-point agreement, an accepted measure for video annotation in the Behavioral Science [50]. A conservative tolerance of one second was set in VData for all variables (two events were said to agree if the secondary coder's mark was within 0.5 seconds on either side of the primary coder's mark).

5.3. A³ Variable Description

The following section is a detailed description of the dependent variables examined in the A³ coding guideline. These descriptions focus on the rationale for each variable and the major choices made when constructing the variable definitions. The actual guide (with the specific topographical or physical features) used is presented in Appendix A. Our annotation process was divided into four passes, each of which asked coders to focus on a specific category of dependent variables while they watched a video in its entirety. Since many of the different variables required different view modes, this pass breakdown not only aided the examination

of the data, but also was optimal for the annotation process. In general, we can divide our dependent variables into those that were based on gross motor behavior and vocal behavior of the subject.

5.3.1. Motor Behavior Variables

With the exception of the metric Time In Chair (Section 5.3.1.5), which was gathered with standard playback, all these metrics were gathered with the Continuous Interval Playback Mode set to three seconds.

5.3.1.1. Smiling

The variable, Smiling, was chosen because it is typically associated with pleasure or enjoyment (e.g. [31]). Although the source of the smile could not always be determined, we hypothesized that we would see a higher rate of smiles during trials the subject enjoyed.

5.3.1.2. No Face

The No Face variable was used to identify three-second intervals when the child's face could not be seen, and no coding determination could be made as to whether or not a smile occurred. This variable was identified because of a concern that surfaced during the coding process; Coders found that when they summarized the data, they had difficulty discerning intervals when no smiles had occurred from those they were unable to code. Although its accuracy is reported, this variable was not directly used in the analysis. Rather its agreement was useful for demonstrating that coders were "on the same page," and allowed agreement calculations for

Smiling, which is dependent upon being able to see the face. (The absence of both Smiling and No Face marks are an indication that the child was not smiling).

5.3.1.3. Oriented at Screen

In order to assess visual attention to content, we created an “orientation arc” for the evaluation of child gaze; see [12, 31] for others who developed this procedure. If gaze was directed within this arc, the subject was considered to be oriented at screen. The arc’s width ($\sim 90^\circ$) was used to accommodate the behavior in which children with autism will use their peripheral vision as primary visual input [41]. See Appendix A for illustration of the orientation arc.

5.3.1.4. Auditory Focus

Much like Oriented to Screen, the Auditory Focus variable was used to assess auditory attention. Unlike visual attention, which has a more observable physical indicator, auditory attention must be observed indirectly. As a result, Auditory Focus was observed via proximity to and physical interaction with the speaker or orientation to the screen/speaker after a new sound was made [28, 69] .

5.3.1.5. Time in Chair

To assess the willingness to attend to computer stimuli, we coded the duration a child would spend in his/her chair [65, 92]. We hypothesized that increased time sitting was a proxy for engagement.

5.3.2. Verbal Variables

Verbal metrics were collected to examine vocalizations during the experimental and control conditions. Coding of vocalizations was facilitated through use of a decision tree, which is incorporated in the full coding guide (Appendix A). With the exception of Turn Taking (Section 5.3.2.7), the following sub-sections are at decision points in the tree rather than one for each variable. These variables were assessed with VCode's Continuous Playback Mode.

5.3.2.1. Child's Sound (*Speech vs. Non-Speech*)

The most basic question coders must address is whether or not a sound is considered "speech-like." Specifically, we define a speech like sound as one that could be phonetically transcribed. This decision point attempts to screen sounds that have the potential to lead to conventional speech and those that may be related to ticks, breathing, self-stimulatory behavior or other forms of expression that are not used in speech production (laughing, screaming, etc) [106].

5.3.2.2. Non-Speech Sounds (*Laughter vs. other*)

Though the range of non-speech sounds is large, we asked coders to distinguish Laughter as another means to examine children's pleasure and engagement during the activity [33]. In addition, we wanted to differentiate Laughter from vocal self-stimulatory behaviors.

Compared to other children with developmental delays, those with ASD tend to produce more non-speech sounds [96]. By annotating non-speech sounds, we also hoped to examine the impact of the external stimuli on their non-speech vocalization.

5.3.2.3. *Speech-Like Sounds (Imitative vs. Spontaneous)*

A critical distinction made in studying the communicative behavior of children with special needs is between sound production that is imitative (repeating a sound previously heard) or spontaneous (without explicit model) [37]. Using this distinction, we hope to explore the types of speech-like sounds produced, and whether there is a direct relationship between what is prompted (human or computer) and what is said.

5.3.2.4. *Imitative Sounds (Immediate vs Differed)*

To explore the imitative sounds produced by subjects, we divided them into those which occur immediately after the source (within five seconds) and those that occur after a more prolonged time [83]. This distinction is particularly relevant for children with autism due to echolalic tendencies [86]. Theory suggests that words/sounds in delayed imitation are stored outside of the subjects short-term or working memory.

5.3.2.5. *Spontaneous Sounds (Orientation to Screen)*

While imitative sounds are, by definition, based on audio stimuli, we wanted to delve deeper into spontaneous sounds, and their relationship to screen orientation. Eye gaze is indicative of engagement and when paired with vocalization, is a key communicative development (e.g [12, 104]). For each spontaneous sound produced, we explored whether that sound was made while oriented to the screen. This allows us to examine the direct relationship between spontaneous sound production and orientation.

5.3.2.6. *Spontaneous Sounds (Immediate vs. Delayed)*

Much like imitative sounds, we wanted to understand if there was any correlation between spontaneous sound production and auditory stimuli. To explore this relationship, we asked coders to mark spontaneous sounds that were made within five seconds (immediate) of a source sound, and those made after a longer period of time (delayed).

5.3.2.7. *Turn Taking*

An important skill in oral communication is that of turn taking, or waiting for others to finish [77, 104]. With all speech-like sounds, we asked coders to determine if the subject waited for the source (be it a researcher or computer generated sound) to “finish” their sound production. In other words, did the child wait for his/her turn to talk (or not interrupt).

5.3.3. *Other Metrics*

There were two other metrics collected not expressly discussed in Section 5.3.1 and 5.3.2; BIGmack Switch and Non-Child Audio.

5.3.3.1. *BIGmack Switch*

The BIGmack™ Switch [1] is an assistive technology device used to play a pre-recorded message for individuals with speech disability [87]. With one subject, who was suspected to have limited motor control of his vocalizations, this device was used to simplify the task of producing speech.

5.3.3.2. Non-Child Audio

These data points served two express purposes. Primarily, they were collected to help clean data logged on the computer by marking sounds (other than those made by the child) in the video that interfered with automatic data gathering. A second purpose was to familiarize coders with the video they were about to watch without forcing them to annotate a very complex variable. Because Non-Child Audio was coded as a first pass, by itself, coders were forced to watch the entire video once, before examining more specific details.

5.4. Resulting Agreement

Overall agreement between coders was good (88%). By the last video, coders spent approximately 20 minutes per one minute of video footage annotating, a significant decrease

Variable	% Agreement
Non child audio	75.96
<i>Duration: Non child audio</i>	82.39
Smiling	90.71
No face	86.01
Oriented at screen	98.09
Auditory focus	87.37
Laugh	63.89
<i>Duration: Laugh</i>	91.30
Non speech vocalization	83.86
Speech like vocalization Sounds	90.50
<i>Duration: Speech like vocalization</i>	92.84
Turn taking	81.60
Immediate + Screen + Spontaneous	79.30
Delayed + Screen + Spontaneous	79.45
Immediate + Spontaneous	78.31
Delayed + Spontaneous	84.62
Differed imitation	<i>None Recorded</i>
Immediate imitation	85.71
Time in chair	88.64
<i>Duration: Time in chair</i>	82.05
BIGMack Switch	87.50

Table 1. Point-by Point Percent Agreement

from the initial 40 minutes per one minute of video footage. One trial from each of the six sessions was randomly selected for agreement checks, totaling six agreement points per child. This represents approximately 11% of the 268 trials from the five subjects. Not all trials were equal in length, nor did all variables occur equally in all sessions. As a result, we examined the agreement values across all randomly sampled sessions. Percent agreements are presented in Table 1. An agreement was defined using a conservative tolerance of one second.

As stated earlier, the point-by-point agreement method was used to calculate our inter-rater

Variable	Kappa
Auditory Focus	0.636
Oriented @ Scrn	0.84
No-Face	0.769
Smiles vs No Smiles	0.63

Table 2. Kappa Statistic for variables coded using interval playback (set to 3 seconds)

agreement value. Because most variables were annotated on an ‘infinite’ timeline (i.e., they could occur an infinite number of times, at an infinite number of locations), we were not concerned about chance agreement. As a result, Kappa [50] calculations do not apply, nor do they make sense when annotating in this way. The statistical Kappa calculation is designed to take into account agreements that occur by chance when calculating inter-rater agreement.

However, for the variables collected using the Continuous Interval Playback (set to 3 seconds) mode, we did calculate a Kappa. A probability of agreement due to chance could be calculated because there was a discrete and finite set of “observation points” (the video was divided into three second intervals and the variables measured in this way were subject to agreement by

chance). The Kappa statistics are presented in Table 2. For Smiling and No Face calculations, we used a 2-tier evaluation metric similar to that used by Reid et al [88]. The first Kappa accounts for the level of observer agreement on whether they could make a judgment about a subject's smiles (could they see the subject's face -- No Face). We then eliminated all of the intervals in which either coder marked No Face implying that they could not assess whether the subject was or was not smiling because the assessment of agreement on Smiling depended on both observers being able to see the subject's face. The remaining intervals, those in which both observers coded the subject as either smiling or not, were examined for agreement. The coding of smiles depended on both observers agreeing that they could see the subject's face.

5.4.1. General Agreement & Accuracy

When taking all of the dependent variables together, the overall inter-rater agreement (IRA) was 88%. Upon closer examination, we determined that 11 of the 21 measured variables had IRA that exceeded 85%, 15 had IRA that exceeded 80%, and 20 exceeding 75%. A further exploration of variables with less than 80% agreement is provided in the following section (5.5.1).

The Kappa statistics calculated from the data suggest a high level of agreement. Kappas ranged from 0.63 (Good) to 0.84 (Very Good). Our interpretation of agreement follows from that of Altman and Byrt [5, 22].

To observe how agreement would change with an increase in the timing tolerance, we increased it from 1.0 second to 5.0 seconds at 0.1 second intervals. Surprisingly, the number of

matched points changed only when the tolerance was increased up to 1.5 seconds (observers' marks were said to agree if the secondary mark was within 0.75 seconds on either side of the primary coder's mark). Of the few variables whose reliability increased, the change resulted in a gain of at most 0.5%. This suggests that the coders were likely accurate in the placement of their marks. Moreover, we can surmise that if there were a lack of agreement in the data, it was likely due to a disagreement of what was coded and not due to ambiguity that an event had occurred.

5.5. Discussion

A short set of questions was posed to our coders, in addition to face-to-face discussions, in an attempt to elicit aspects of the coding guidelines that were difficult or that they believed to be overly subjective. Their feedback, in addition to our observations, are discussed below in relation to the agreement results. We also discuss the impact of the A³ system on the outcomes of the original study for which it was the system of measurement.

5.5.1. Difficulties in Agreement

Three distinct groups/variables emerged for which 80% agreement was not reached: Laughter, Non-Child Audio and Spontaneous Speech Like Vocalization Variables. We discuss these areas, coder's feedback, and their implications for the A³ system.

5.5.1.1. Laughter

The Laughter variable had the lowest agreement values (63.89%) of all the dependent variables collected. However, upon further inspection, we noticed that it also had a low frequency of

occurrence. It is well known in research that employs observational annotation that the lowest levels of agreement are achieved for low rate behavior [18].

Variable	Agreement
Immediate Spontaneous	83.2%
Delayed Spontaneous	80.6%
All Spontaneous	84.6%

Table 3. Combined data from the four spontaneous speech like vocalization variables

Coders also mentioned difficulty in distinguishing Laughter from Speech-Like Vocalizations and Non-Speech Vocalizations. Often, coders found that vocalizations may have been laughter-like but matching positive affect display with the vocalization was difficult to pinpoint. Perhaps this difficulty is exacerbated by the differences in affective expression that characterizes ASD [104].

5.5.1.2. Spontaneous Speech-Like Vocalizations

The sub-divisions within Spontaneous Speech-Like Vocalizations also resulted in low agreement between coders (78.31% to 84.62% with a mean of 80.42%). To explore the effect of sub-division on reliability, we combined data across variables (Table 3). To combine two variables, we treated all marks for both variables the same, and re-calculated agreement. This analysis can suggest at what level of granularity (distinction between variations of a variable) these variables can be reliably coded. As we combined, we noticed an improvement in agreement reaching 80% once we eliminated the distinction between sounds made while looking at the screen, versus those made without. It appears that such small distinctions between sounds made while looking at the screen and those when looking away may have been

too fine-grained a distinction to accurately code. For data analysis, we intend to only differentiate Immediate and Delayed Spontaneous Speech.

Upon further examination of the Spontaneous Speech-Like data, we discovered the potential for double-counting disagreements. Every Speech-Like Vocalization was coded as either Spontaneous or Imitative. However, every disagreement in Speech-Like Vocalization is a guaranteed disagreement for the Spontaneous/Imitative distinction. Thus, our lower agreement values may have been a direct result of “double counting.”

5.5.1.3. Non-Child Audio

The second variable that had less than 80% agreement was Non-Child Audio. This variable, defined as any sound not produced by the child, appeared to cause difficulties due to the quality of the audio recording. Some coders were better able to hear “quieter” sounds and, thus, would mark them. As a result, there was a discrepancy between the coders. We propose an audio “threshold” be set to differentiate between sounds to code and sounds not to code. This differentiation could be visualized in VCode for ease of the coders. In the end, however, Non-Child Audio was not used in the study.

5.5.2. Other Feedback & Observations

The majority of the time annotating was spent determining if Speech-Like Sounds were Imitative or Spontaneous. Although coders mentioned that they found this differentiation frustrating, they also found no way to improve the variable definitions to increase accuracy or speed. We believe that this frustration stemmed from audio quality. Poor audio, in

conjunction with a population that generally has poor articulation, may have resulted in difficulty differentiating specific sounds. Similarly, at the start of this research, we had hoped to transcribe phonetically the sounds made by the children to examine their phonetic repertoires. However, we also discovered that audio quality was crucial for this form of transcription. We eliminated coding of the phonetic repertoire, although we hope to reintroduce it in the future in conjunction with better audio recording techniques.

In discussion, coders mentioned some confusion over the No Face variable, specifically in the boundary condition when the child has part of his or her face covered. Feedback from coders included specific requests for a more explicit definition of the features that must be seen to justify annotating No-Face. For future experiments, we propose specifying features of the face (e.g., lips, cheeks) necessary to determine whether or not a child is smiling.

One other point of improvement suggested by the coders was the definition of Smiling. Coders asked that future guides specify the exact facial features that did or did not constitute a smile.

5.6. Forms of Use

Though validation of A³ came in the context of coding all variables for video of children with ASD interacting with technology, we strongly believe that the application of A³ goes well beyond this particular situation.

5.6.1. Research External to ASD

Because this guideline focuses on non-verbal subjects interacting with technology, we believe that A³ can be applied to other areas of HCI research that target non-verbal subjects. This can include infants, subjects with Apraxia, or other vocal impairments. Because our system also covers focus on computer feedback and categories of vocalizations, A³ could also be used in situations where subjects use speech devices (e.g. [6]).

5.6.2. Use of the A³ Guidelines

Though A³ does not cover every possible dependent variable for video analysis, it does provide a robust library of features to annotate. However, not all features may be applicable, or worth analyzing for every experiment. As a result, A³ can be used as a source for dependent variable selection. Researchers can select the most applicable set and definition of variables from the guideline, and use that sub set in their own research.

Further, variables do not simply need to be used by occurrence count. Frequency of occurrence (count/video duration) and ratio analysis (variable A: variable B) allows for a more robust set of methods for behavior analysis. We believe that the strength of A³ is it's flexibility to allow researchers to examine their own research in the most applicable way, using a set of demonstrably codable dependent variables.

5.7. Summary

ASD-related tools create new challenges to software developers, due to the different subject demographic compared to existing techniques from other forms of assistive technology.

Because no such coding scheme exists for ASD-related tools, it is incumbent on us to faithfully describe our scheme and its relationship to other available tools, and to provide data to support its reliability in the field. We proposed a new set of dependent variables to be assessed through the video annotation process called A³ (Annotation for ASD Analysis). This set has been tested in a research context and the data collected have produced meaningful results about the behavior of low-functioning children with ASD and their interaction with audio and visual feedback systems. The chapter demonstrated A³'s reliability, and discussed shortcomings and areas of improvement for the coding guides.

With this set of dependent variables, we have operationalized the coding process through detailed descriptions of the dependent variables and use of the VCode and VData system. As a result, time for annotation has been reduced to 20 minutes per 1 minute of footage, while still maintaining adequate reliability. The research and development of VCode, VData and A³, allow us to assess the interaction of our subjects and SIP. The analysis of this data is described in the following chapters.

CHAPTER 6: DATA ANALYSIS⁶

Over a six-month period, 1200 minutes of video were annotated with all variables in the A³ guidelines. However, we focused our analysis on Spontaneous Speech-Like Vocalizations (SSLV), one of the dependent variables from A³. There is clear and important distinction between those vocalizations that are spontaneous and those that are imitative. This is critical when assessing children with special needs [37].

Spontaneous Speech-Like Vocalizations (SSLV)— sounds produced by the subject that could 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.

6.1. 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. This facilitated contrast between the mode of feedback (visual, auditory, and mixed) as well as the different types of feedback (12 visual and 5 auditory forms).

⁶ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

6.2. Question Analysis

Each subject was analyzed separately. Due to the varying lengths of each trial, a comparison between the number of occurrences of SSLV would be weighted towards longer sessions. To mitigate this effect, we analyzed a normalized frequency of SSLV (occurrences in trial divided by trial duration). Wilcoxon rank-sum and Kruskal-Wallis tests were used to compare the number of SSLV 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.

6.2.1. Q1 Analysis

Q1 examines if there is at least one form of computer generated feedback that will positively impact a subject's frequency of SSLV. If there at least one condition in Q2-Q4 shows feedback has a positive impact on frequency of SSLV, we can conclude R1 is true for that subject.

6.2.2. Q2 Analysis

Q2 indicates, in general, that all forms of feedback (regardless of mode/style) increase frequency of SSLV. Analysis of Q2 for each subject is determined by comparing the frequency

of SSLV at baseline to frequency across all types of feedback using the Wilcoxon rank-sum test.

6.2.3. *Q3 Analysis*

Q3 indicates if all forms of feedback in a specific modality positively impact SSLV. Analysis of Q3 for each subject is determined by performing a Wilcoxon rank-sum test comparing frequency of SSLV at baseline with frequency of SSLV in groups audio only, video only, and mixed feedback. Results from Q3 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 have a significant p value, Q3 is considered Neither, indicating that no modality increased the frequency of SSLV (all $p > 0.05$).

6.2.4. *Q3a Analysis*

Q3a examines if there is a specific type of feedback that increased frequency of SSLV in a modality that approached significance. Using the result from Q3, we will tease out specific forms/combinations of feedback within those statistically significant modalities (visual, auditory, mixed). Trials within the specific modality are broken down into subcategories based specific forms of feedback and tested against baseline using the Wilcoxon rank-sum test. Q3a is only asked if p values for Q3 were approaching statistical significance.

6.2.5. Q4 Analysis

We used qualitative observations from researchers and video to further guide analysis. This enabled us to utilize overlooked forms of feedback that increased frequency of SSLV. Using the Wilcoxon rank-sum test, we compared baseline with conditions that were qualitatively observed to increase SSLV frequency.

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

6.2.6. Q5 Analysis

In order to categorize the forms of feedback which illicit an increase in SSLV frequency, we extracted the mode of feedback found to have the most impact in Q3, Q3a and Q4. This synthesis of results provides a better understanding of what modes of feedback are engaging.

Using these questions as guides, we analyzed the results from each of our subjects to assess the impact of computer based visual and auditor feedback on SSLV. The following chapter describes the analysis in detail by presenting the findings broken up by subject.

CHAPTER 7: RESULTS & DISCUSSION⁷

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 [79] were in the first phase (Preverbal Communication), roughly equating to the development of a neurologically typical 6-12 month old.

	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]

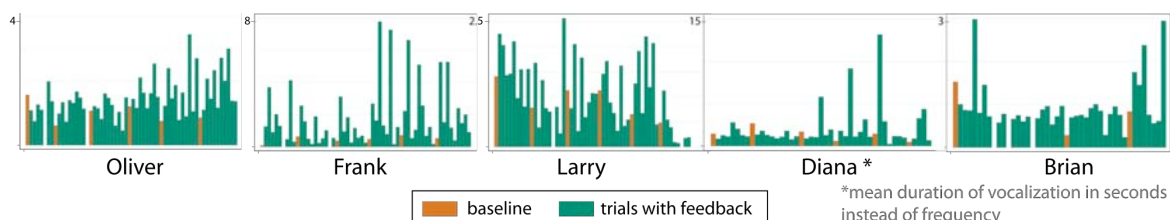


Figure 12. Demographics and Frequency of SSLV: Wilcoxon Rank-Sum Test from R2 and R3 Analysis.
High level graphical comparison of Frequency of SSLV per 10 seconds across all trials for all subjects.

⁷ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

7.1. Subject 1: Oliver

7.1.1. Oliver's Results

Initial analysis of Oliver's data (Figure 12) demonstrated borderline significance comparing baseline to all feedback (Q2). Further, the audio only and mixed feedback conditions (R4) approach significance. Due to a trend towards significance in the two conditions involving audio, we compared frequency 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 conclude that audio feedback may have played a role in increasing the

	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]	0.042 [-2.03]

Table 4. Comparison of Oliver's audio feedback

frequency of Oliver's SSLVs (Q5).

Since audio appeared to increase the frequency of Oliver's SSLVs, we explored impact of different forms of audio feedback in combination with visual feedback. Table 4 shows that echo feedback encouraged SSLV, while visual feedback did not appear to have significant impact on SSLV frequency (Q3a). We qualitatively observed that Oliver reacted positively to

audio from a popular cartoon show. Our data confirms this by approaching statistical significance ($p=0.083 [-1.74]$) (Q4).

From this analysis, we conclude that Oliver increased his frequency of SSLV in conditions with audio feedback. Specifically, he increased SSLV in conditions with echoing audio feedback (Q1).

7.2. Subject 2: Frank

7.2.1. Frank's Results

Initial analysis of Frank's data (Figure 12) showed a significant difference in frequency of baseline SSLVs and frequency of SSLVs with all feedback (Q2). We found a statistically significant difference in frequency of SSLVs with audio only and mixed feedback (Q3). Due to

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 5. Comparison of Frank's audio feedback

significance in both conditions with audio, we compared frequency of baseline SSLVs with any condition with audio feedback. There was a highly significant association between audio feedback and SSLVs ($p = 0.004 [-2.84]$) (Q5).

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

Table 6. Frank: Form of visual feedback with any audio

Given the robust effect of audio feedback, we compared Frank's responsiveness to audio feedback with and without visual feedback (Table 5). 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 frequency of SSLVs over baseline when audio from the cartoon was present (Q3a, Q4). For this subject, visual feedback had a positive impact on the frequency of SSLVs when audio was also present.

Finally, we examined all conditions with audio feedback into specific forms of visual feedback to assess the impact of different forms of visual feedback on the frequency of SSLV production. Based on qualitative observations, we analyzed trials where a visual image from a specific cartoon was present. Frank demonstrated increased SSLV frequency over baseline for all visual feedback in addition to audio for all but Spinning Spiral of Dots and Random Dots (Table 6), with the highest significance in Firework-Like Feedback (Q3a).

From this analysis, we conclude that Frank had a higher frequency of SSLV to conditions with audio feedback and both audio and visual feedback together (Q1 , Q5). Specifically, he appeared to show increased SSLV when audio and visuals from a specific cartoon.

Interestingly, his mother stated that Frank did not watch this cartoon show.

7.3. Subject: Larry

7.3.1. Larry's Results

Initial analysis of Larry's data (Figure 12) failed to reach statistical significance (Q2, Q3).

While formal statistical tests did not reach statistical significance, qualitative observations

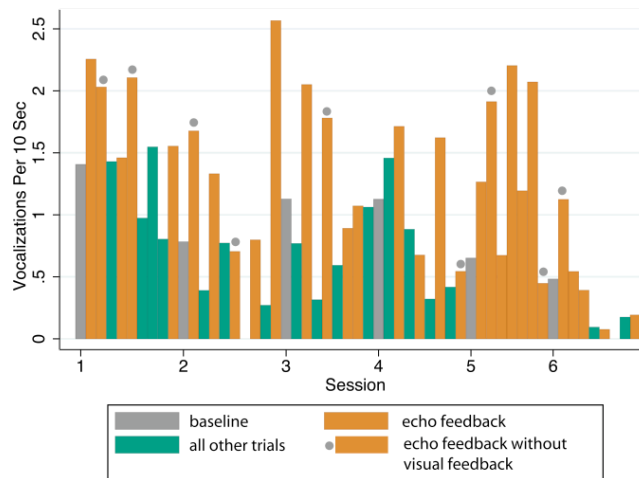


Figure 13. Larry's SSLV frequency, by session and trial.

from researchers and study video, in conjunction with graphical representation of the data (Figure 13) led us to believe that there was feedback that had impact on frequency 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 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.

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

To compare the impact of echoing feedback on SSLV with other forms 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) only audio feedback (excluding echoing), (4) only visual feedback (excluding echoing), (5) audio + visual feedback (excluding echoing). The Kruskal-Wallis test had a $p=0.060$. To increase statistical power, we collapsed groups⁸ by combining visual only feedback with mixed condition since groups had visual presentations (comparative analysis between collapsed groups: Wilcoxon rank-sum

⁸ Collapsing two groups increases the number of data points in the resulting group, thus increasing the statistical power during comparison.

$p=1.000[0.00]$). 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 7) was performed. Statistically significant differences were found between the echo condition and audio only (visual + mixed) ($p=0.034$, $p=0.023$ respectively) (Q4).

From this analysis, we conclude that Larry showed preference for echoing audio feedback (Q1, Q5). However, we believe that with more statistical power, we could make a more conclusive statement.

7.4. Subject: Diana

Diana responded to many commands by her mother such as sit, stop, come here, and wait.

Diana demonstrated two to three signs for communication (e.g., more, music), though articulation of signs was poor, and frequency was low (about 1 per session).

7.4.1. Diana's Results

Initial data analysis for Diana, found much higher p values (0.5-0.9) than expected when comparing them to qualitative notes made by researchers. Confused by these findings, we examined annotations made by video coders and noticed that large strings of Diana's SSLVs were being grouped together. A³ 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 just one occurrence. To accommodate her shorter pauses, we re-analyzed her data using mean duration

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

of SSLVs rather than frequency. For this subject, we used average duration as a proxy for frequency.

Initial analysis of duration of SSLV (Figure 12) showed significance for visual only conditions (Q2, Q3). Audio only feedback was not used, due to lack of interest observed in initial orientation sessions.

To examine impact of visual feedback, we broke down the forms of visual only feedback and compared average duration of spontaneous SSLVs with those produced in baseline condition (Table 8). The last row in Table 5 is an amalgam of different forms of visual feedback in which abstract colored dots are replaced with one or more found image(s). This data support our qualitative observations that Diana only responded to conditions where images shown were from cartoon shows, and that audio feedback reduced her SSLV (Q4). Three statistically significant conditions were Spinning Image (a found image from a cartoon spins on axis),

Multiple Circles (many dots or found images appear on screen; size based on volume of sound produced) and any feedback with Found Images (there are overlaps between groups) (Q3a, Q4).

From this analysis, we conclude that Diana produced more SSLVs (mean duration) with visual feedback compared to baseline and mixed (Q1, Q5). Specifically, she appeared to show increased engagement with forms of visual feedback that contained a cartoon character (though a specific preference did not appear). Diana was reported to watch movies/TV-shows with these characters.

7.5. Subject: Brian

7.5.1. 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 on the other subjects (Figure 12). 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.

From this analysis, we could not conclude that Brian had a significant reaction to any form of feedback (either compared to baseline or against each other) (Q1-Q5).

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

7.6. Discussion

After a thorough examination of the quantitative data collected, we are able to summarize the findings in relation to our 5 questions (Table 9).

7.6.1. Q1 Discussion

In 4 of the 5 subjects, we found that at least one form of feedback created an increased frequency of SSLVs. We were unable to show that any form or modality of feedback, when compared to baseline, significantly increased the frequency of SSLVs 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 frequency of SSLVs when compared with all other forms of feedback for Larry. Overall, we conclude that feedback may encourage SSLV in children with ASD.

7.6.2. Q2 Discussion

Only one of five subjects found all forms of feedback, regardless of mode or form, to have a positive impact on frequency of SSLV. This finding suggests that not all forms of computer feedback work for all children.

7.6.3. Q3 & Q5 Discussion

It is commonly believed that individuals with ASD respond better to visual feedback than auditory [8, 34, 73]. 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). One subject (Brian) did not show any significant reaction to any form of feedback. When taken from a more global level, 3 of 5 subjects responded to audio feedback, and 2 of 5 responded to visual feedback Table 9. This suggests that further exploration of feedback in both visual and audio modality is essential. This finding is of particular note in that it is in contrast to other work.

7.6.4. Q3a & Q4 Discussion

Though some subjects had a larger range of forms of feedback that resulted in increased frequency of SSLV than others, 4 of 5 subjects did have one particular condition that outperformed the others. The specific results, in conjunction with varied modes of feedback that resulted from Q3 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 forms of audio/visual feedback that garnered the increase in SSLV.

7.6.5. Parental Response

In addition to data from subjects during the sessions, we asked for 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 solutions to be put to use 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.

CHAPTER 8: FOLLOW UP STUDY⁹

8.1. Wizard-of-Oz Follow Up Study

One future application of the SIP framework, is to encourage specific forms of vocalization, namely teaching vocabulary acquisition and pronunciation. Given the current state of SIPS, this particular functionality is not readily available. However, a common technique used in design is a “Wizard-of-Oz” study. This is a form of experimentation in which a researcher operates a partially functional system to make it appear to be fully functional to subjects.

Researchers constructed a Wizard-of-Oz system, based on SIP, geared towards teaching specific skills. The model followed a common form of Behavioral Therapy [64]: Prompt for word – wait for response – reward if correct or repeat if incorrect. We replaced the computer voice recognition with a researcher to test the concept. Researchers qualitatively noted Frank’s response as being exceptional, both in terms of his reaction to the computer feedback and his eagerness to participate. Noting this, researchers asked his mother’s permission to include him in an additional Wizard-of-Oz study.

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

⁹ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

instantaneous reaction to their sounds, for both visual and auditory reinterpretation. If the 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 repeated every prompted word.

During the second session, we used 6 words his mother stated 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 repeat the new words. Though articulation was often unclear he made a concerted effort to repeat all 10 words, including the 6 new ones. Of particular note, Frank has been highly resistant in the past with this form of Vocal Imitation Language therapy.

CHAPTER 9: CONCLUSION¹⁰

This thesis presents the motivation, background and theory related to the use of visually and auditory feedback to encourage vocalization in non-verbal children with ASD. Based on the existing literature, we developed SIP or the Spoken Impact Project. SIP consists of three major thrusts of research; 1) A framework to facilitate video annotation, and an instantiation of said system called VCode and VData; 2) A³, a set of coding guidelines that allow researchers to assess non-verbal subject interaction with computer based feedback; and 3) A detailed quantitative analysis of SIPS through the use of VCode, VData and A³. Results from this analysis suggest that computer based feedback systems hold the potential to greatly impact the vocalization of non-verbal subjects. Based on our follow up Wizard of Oz Study, we believe that systems like SIPS can be further developed to not only encourage vocalization, but also teach verbal communication skills.

9.1. 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 form of feedback, while other forms 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 our trials had high variance, and reduction in observation

¹⁰ Some research, content, and text from this section is in submission to ACM under the title: *Creating a Spoken Impact: encouraging vocalization through audio visual feedback in children with ASD* with authors Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. S. and Coletto, M. This content is printed with knowledge of the Association for Computing Machinery.

time may have reduced statistical power of this study and ability 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 forms of feedback that garnered significant changes in SSLV.

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 single-subject studies, 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. This work, in conjunction with the findings from our Wizard-Of-Oz study, lay the ground work for future exploration of this area of research.

9.2. Future Work

Given our encouraging results, there are many exciting areas of future work. One of the most immediate directions is adaptive feedback selection. Previously, researchers had to qualitatively assess which visualizations and forms of audio feedback were engaging to subjects. Future work might examine if a system could adaptively change forms of feedback by 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 be able to adapt.

We see the potential to test our approach with other populations or other target behaviors. One unanswered question is the 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 the practitioner's office.

9.3. Summary

The implications of this thesis are broad and far reaching. Results from our research on video annotation provide a framework for software designers to better meet the needs of researchers, clinicians, and to improve video annotation accuracy and quality. In addition, two tools that embody the annotation framework have been rereleased, with impact reaching world wide to a plethora of disciplines from Neuroscience to Computer Science.

The A³ guidelines demonstrate how research from multiple domains can be brought together to create a system for analyzing the behavior of non-verbal subjects. By leveraging the rich body of literature, we were able to build a robust framework that leverages the theory and years of research already conducted. Based on the findings for inter coder reliability, the A³ guidelines provide a reliable system to quantitatively assess subject behavior. A³ used in conjunction with VCode and VData create a system that is reliable, fast, and accurate for assessing non verbal subject behavior.

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 suggests 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 that states the strong preference for visual only feedback.

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

With the positive 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 the first steps towards uncovering the area of using audio and visual feedback to encourage speech in low functioning children with autism. In other words, SIP is a starting point for future research.

References

- [1] AbleNet®. BIGmack® communicator - Red.
http://www.ablenetinc.com/item_detail.aspx?ItemCode=10000201 Roseville, MN,
Accessed 2008
- [2] Abowd, G. D., Gauger, M. and Lachenmann, A. The Family Video Archive: an annotation and browsing environment for home movies. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval* (Berkeley, California, 2003). ACM, 2003.
- [3] 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.
- [4] Ahn, H. and Picard, R. *Affective cognitive learning and decision making: The role of emotions*. City, 2006.
- [5] Altman, D. *Practical Statistics/or Medical Research*. Chapman and Hall, London, 1991.
- [6] Attainment Company GoTalk. City, 2008.
- [7] Autism Society of America, A. Autism Society of America. <http://www.autism-society.org> Accessed 2007
- [8] Baggs, A. *In My Language*. YouTube, City, 2007.
- [9] Banerjee, S., Cohen, J., Quisel, T., Chan, A., Patodia, Y., Al-Bawab, Z., Zhang, R., Rybski, P., Veloso, M. and Black, A. Creating multi-modal, user-centric records of meetings with the Carnegie Mellon meeting recorder architecture. In *Proceedings of the ICASSP Meeting Recognition Workshop* (Montreal, Quebec, Canada, 2004), 2004.
- [10] 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.
- [11] Barry, R. M. EPG from square one: An overview of electropalatography as an aid to therapy. *Clinical Linguistics & Phonetics*, 3, 1 (1989), 81-91.
- [12] 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.
- [13] Bergstrom, T. and Karahalios, K. Conversation Clock: Visualizing Audio Patterns in Co-Located Groups. In *Proceedings of the HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCE 2007* (Hawaii, USA, 2007). IEEE, 2007.

- [14] Bergstrom, T. and Karahalios, K. Conversation votes: enabling anonymous cues. In *Proceedings of the CHI '07 extended abstracts on Human factors in computing systems* (San Jose, CA, USA, 2007). ACM, 1240994
2279-2284, 2007.
- [15] Bergstrom, T. and Karahalios, K. Seeing More: Visualizing Audio Cues. In *Proceedings of the INTERACT* (Rio de Janeiro, Brasil, 2007). ACM Press., New York, NY, 2007.
- [16] Berry, K. and Mielke Jr, P. A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters. *Educational and Psychological Measurement*, 48, 4 (1988), 921.
- [17] Bertini, M., Del Bimbo, A., Cucchiara, R. and Prati, A. Semantic video adaptation based on automatic annotation of sport videos. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval* (New York, NY, 2004). ACM New York, NY, USA, 2004.
- [18] Birkimer, J. C. and Brown, J. H. A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis*, 12, 4 (Winter 1979), 523-533.
- [19] Bondy, A. and Frost, L. The Picture Exchange Communication System. *Behavior Modification*, 25, 5 (2001), 725-744.
- [20] Bonnef, Y. S., Belmonte, M. K., Pei, F., Iversen, P. E., Kenet, T., Akshoomoff, N., Adini, Y., Simon, H. J., Moore, C. I., Houde, J. F. and Merzenich, M. M. Cross-modal extinction in a boy with severely autistic behaviour and high verbal intelligence. *Cognitive Neuropsychology*, 25, 5 (2008), 635 - 652.
- [21] Burr, B. VACA: a tool for qualitative video analysis. In *Proceedings of the CHI '06 extended abstracts on Human factors in computing systems* (Montreal, Quebec, Canada, 2006). ACM, 2006.
- [22] Byrt, T. How good is that agreement? *Epidemiology*, 7, 5 (September 1996), 561.
- [23] 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.
- [24] Cassell, J., Kopp, S., Tepper, P., Ferriman, K. and Striegnitz, K. Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions. John Wiley & Sons, New York, 2007.
- [25] Center for Disease Control and Prevention, C. Autism Information Center, DD, NCBDDD, CDC. <http://www.cdc.gov/ncbddd/autism/> Atlanta, Accessed 2007
- [26] Chen, S., Shyu, M., Liao, W. and Zhang, C. *Scene change detection by audio and video clues*. City, 2002.

- [27] Clifford, J., Marcus, G. E. and School of American Research *Writing Culture*. University of California Press, Berkeley, California, 1986.
- [28] Clifton, R. K., Perrisa, E. E. and McCalla, D. D. Does reaching in the dark for unseen objects reflect representation in infants? *Infant Behavior and Development*, 22, 3 (1999), 297-302.
- [29] Conger, A. Kappa Reliabilities for Continuous Behaviors and Events. *Educational and Psychological Measurement*, 45, 4 (1985), 861.
- [30] Costa, M., Correia, N., Guimar, N. and es. Annotations as multiple perspectives of video content. In *Proceedings of the ACM international conference on Multimedia* (Juan-les-Pins, France, 2002). ACM, 2002.
- [31] Field, T., Field, T., Sanders, C. and Nadel, J. Children with Autism Display more Social Behaviors after Repeated Imitation Sessions. *Autism*, 5, 3 (Sep 2001), 317-323.
- [32] Fry, B. and Reas, C. *Processing*. 2007
- [33] Gena, A., Krantz, P., McClannahan, L. and Poulson, C. Training and Generalization of Affective Behavior Displayed by Youth with Autism. *Journal of Applied Behavior Analysis*, 29, 3 (Fall 1996), 291-304.
- [34] Grandin, T. *Thinking in Pictures: And Other Reports from My Life with Autism*. Vintage Books, New York, 2006.
- [35] 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).
- [36] 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 the ACM SIGACCESS- ASSETS 2008* (Halifax, Canada, 2008). ACM-PRESS, New York, NY, 2008.
- [37] Halle, J. Teaching Language in the Natural Environment: An Analysis of Spontaneity. *Journal of the Association for Persons with Severe Handicaps (JASH)*, 12, 1 (Spring 1987), 28-37.
- [38] Hayes, G. R., Kientz, J. A., Truong, K. N., White, D. R., Abowd, G. D. and Perin, T. Designing Capture Applications to Support the Education of Children with Autism In *Proceedings of the International Conference on Ubiquitous Computing* (Nottingham, England, 2004), 2004.
- [39] Hayne, H., Gross, J., Hildreth, K. and Rovee-Collier, C. Repeated reminders increase the speed of memory retrieval by 3-month-old infants. *Developmental Science*, 3, 3 (August 2000), 312-318.

- [40] Heer, J., Viégas, F. B. and Wattenberg, M. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (San Jose, California, USA, 2007). ACM, 2007.
- [41] Howlin, P. An Overview of Social Behavior in Autism. Plenum, New York, NY, 1986.
- [42] IBM. *Speech Viewer III*. 1997
- [43] Johnson, A. About vprism video data analysis software.
<http://www.camse.org/andy/VP/vprism.htm> Accessed 2007
- [44] Kanner, L. Autistic Disturbances of Affective Contact. V.H. Winston, 1943.
- [45] Karahalios, K. and Donath, J. Telemurals: linking remote spaces with social catalysts. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (Vienna, Austria, 2004). ACM New York, NY, USA, 2004.
- [46] Karahalios, K. and Viégas, F. B. VisiPhone. In *Proceedings of the ACM SIGGRAPH 99 Conference abstracts and applications* (Los Angeles, California, United States, 1999). ACM, 1999.
- [47] Karahalios, K. G. and Dobson, K. Chit chat club: bridging virtual and physical space for social interaction. In *Proceedings of the CHI '05 extended abstracts on Human factors in computing systems* (Portland, OR, USA, 2005). ACM, 2005.
- [48] Karahalios, K. G. and Viégas, F. B. Social visualization: exploring text, audio, and video interaction. In *Proceedings of the CHI '06 extended abstracts on Human factors in computing systems* (Montréal, Québec, Canada, 2006). ACM, 2006.
- [49] KayPentax. *Visi-Pitch IV, Model 2950B*. 1996-2008
- [50] Kazdin, A. E. *Single-Case Research Designs: Methods for Clinical and Applied Setting*. Oxford University Press, USA, 1982.
- [51] 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 the fifth international ACM conference on Assistive technologies* (Edinburgh, Scotland, 2002, 2002). ACM Press, New York, NY, 2002.
- [52] 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 the SIGCHI conference on Human factors in computing systems* (San Jose, California, USA, 2007). ACM Press, New York, NY, 2007.
- [53] 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 the 2006 20th anniversary conference on Computer supported cooperative work* (Banff, Alberta, Canada, 2006). ACM Press, New York, NY, 2006.

- [54] Kipp, M. *Anvil - the video annotation research tool*. 2007
- [55] 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.
- [56] Koegel, R. L., Camarata, S., Koegel, L. K., Ben-Tall, A. and Smith, A. E. Increasing Speech Intelligibility in Children with Autism. *Journal of Autism and Developmental Disorders*, 28, 3 (June 1998), 241-251.
- [57] Leadholm, B., Miller, J. and Children, B. f. E. *Language Sample Analysis: The Wisconsin Guide*. Wisconsin Dept. of Public Instruction, 1992.
- [58] Lee, L. *Developmental Sentence Analysis*. Northwestern University Press, Evanston, IL, 1974.
- [59] Lehman, J. F. Toward the use of speech and natural language technology in intervention for a language-disordered population. In *Proceedings of the Proceedings of the third international ACM conference on Assistive technologies* (Marina del Rey, California, United States, 1998). ACM Press, New York, NY, 1998.
- [60] Leonard, L. B. *Children with Specific Language Impairment*. MIT Press, Cambridge, MA, 2000.
- [61] Leung, S. T. Integrating visualization to make programming concepts concrete: dot net style. In *Proceedings of the Proceedings of the 7th conference on Information technology education* (Minneapolis, Minnesota, USA, 2006). ACM, 1168848
149-156, 2006.
- [62] Levin, G. and Lieberman, Z. In-situ speech visualization in real-time interactive installation and performance. In *Proceedings of the Non-Photorealistic Animation and Rendering* (Annecy, France, 2004). ACM, 2004.
- [63] Lord, C., Risi, S., Lambrecht, L., Cook, E. J., Leventhal, B., DiLavore, P., Pickles, A. and Rutter, M. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30, 3 (June 2000), 205-223.
- [64] Lovaas, I. I. *The Autistic Child*. John Wiley & Sons, Inc, New York, 1977.
- [65] Lovaas, O. I. *Teaching Individuals with Developmental Delays: Basic Intervention Techniques*. PRO-ED, Inc., Austin, TX, 2003.
- [66] Luo, Y. and Baillargeon, R. Can a Self-Propelled Box Have a Goal? *Psychological Science*, 16, 8 (2005 2005), 601-608.
- [67] Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S. and Ames, M. Heuristic evaluation of ambient displays. In *Proceedings of the Proceedings of the SIGCHI conference*

- on *Human factors in computing systems* (Ft. Lauderdale, Florida, USA, 2003). ACM, 2003.
- [68] Marshalla, P. *Becoming Verbal With Childhood Apraxia: New Insights on Piaget for Today's Therapy*. Marshalla Speech and Language, Kirkland, WA, 2001.
 - [69] McCalla, D. D. and Clifton, R. K. Infants' means-end search for hidden objects in the absence of visual feedback. *Infant Behavior and Development*, 22, 2 (January 1999), 179-195.
 - [70] Michaud, F. and Théberge-Turmel, C. *Mobile robotic toys and autism*. Springer, 2002.
 - [71] 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.
 - [72] Mohamed, A. O., Courboulay, V., Sehaba, K. and Menard, M. Attention analysis in interactive software for children with autism. In *Proceedings of the ACM SIGACCESS Conference on Assistive Technologies* (Portland, Oregon, October 23 - 25, 2006). ACM, 2006.
 - [73] Mukhopadhyay, T. R. *Beyond the Silence: My Life, the World and Autism*. National Autistic Society, London, 2000.
 - [74] Nakagawa, S. A Survey on Automatic Speech Recognition. *IEICE TRANSACTIONS on Information and Systems*, E85-D, 3 2002), 465-486.
 - [75] National Research Council. *Educating Children with Autism*. Division of Behavioral and Social Sciences and Education, Washington, DC: National Academy Press, 2001.
 - [76] Noldus Information Technology. *Noldus. The observer*. 2007
 - [77] Owens, R. E. *Language Development: An Introduction (7th Edition)* Allyn & Bacon, Boston, MA, 2007.
 - [78] 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 the Proceeding of the 2005 conference on Interaction design and children* (Boulder, Colorado, 2005). ACM Press, New York, NY, 2005.
 - [79] Paul, R. Recommendation for Benchmarks - Autism Speaks Luncheon at SRCLD. L. DeThorne. 2008.
 - [80] 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.
 - [81] Plaue, C., Miller, T. and Stasko, J. Is a picture worth a thousand words?: an evaluation of information awareness displays. In *Proceedings of the Proceedings of Graphics Interface 2004*

- (London, Ontario, Canada, 2004). Canadian Human-Computer Communications Society, 2004.
- [82] Poncelion, D. and Srinivasan, S. Automatic discovery of salient segments in imperfect speech transcripts. In *Proceedings of the tenth international conference on Information and knowledge management* (Atlanta, Georgia, 2001). ACM New York, NY, USA, 2001.
 - [83] Prizant, B. M., Schuler, A. L., Wetherby, A. M. and Rydell, P. Enhancing language and communication: Language approaches. Wiley, New York, 1997.
 - [84] Quek, F., McNeill, D., Rose, T. and Shi, Y. A coding tool for multimodal analysis of meeting video. In *Proceedings of the NIST Meeting Room Workshop* (Montreal, Quebec, Canada, 2003), 2003.
 - [85] Ramos, G. and Balakrishnan, R. Fluid interaction techniques for the control and annotation of digital video. In *Proceedings of the 16th annual ACM symposium on User interface software and technology* (Vancouver, Canada, 2003). ACM, 2003.
 - [86] Rapin, I. and Dunn, M. Language disorders in children with autism. *Seminars in Pediatric Neurology*, 4, 2 (June 1997), 86-92
 - [87] Reichle, J., Beukelman, D. and Light, J. *Implementing an augmentative communication system: Exemplary strategies for beginning communicators*. Brookes Publishing Company, Baltimore, MD, 2002.
 - [88] Reid, D. H., Parsons, M. B., McCarn, J. E., Green, C. W., Phillips, J. F. and Schepis, M. M. Providing a more appropriate education for severely handicapped persons: increasing and validating functional classroom tasks. *Journal of Applied Behavior Analysis*, 18, 4 (Winter 1985), 289-301.
 - [89] Retherford, K., Sowards, D. and Hess, L. *Guide to Analysis of Language Transcripts*. Thinking Publications Eau Claire, WI, 1993.
 - [90] Rosenblum, K., Zeanah, C., McDonough, S. and Muzik, M. Video-taped coding of working model of the child interviews: a viable and useful alternative to verbatim transcripts? *Infant Behavior and Development*, 27, 4 (2004), 544-549.
 - [91] Russo, N., Larson, C. and Kraus, N. Audio-vocal system regulation in children with autism spectrum disorders. *Experimental Brain Research*, Volume 188, 1 (June 2008 2008).
 - [92] Sajwaj, T., Twardosz, S. and Burke, M. Side effects of extinction procedures in a remedial preschool. *Journal of Applied Behavior Analysis*, 5, 2 (Summer 1972), 163-175.
 - [93] SALT Software LLC. *Salt Software*. 2007
 - [94] SaySoft. Annotation. <http://www.saysosoft.com/> Accessed 2007

- [95] Segal, L. B., Oster, H., Cohen, M., Caspi, B., Myers, M. and Brown, D. Smiling and Fussing in Seven-Month-Old Preterm and Full-Term Black Infants in the Still-Face Situation. *Child Development*, 66, 6 1995), 1829-1843.
- [96] Sheinkopf, S. J., Mundy, P., Oller, D. K. and Steffens, M. Vocal Atypicalities of Preverbal Autistic Children. *Journal of Autism and Developmental Disorders*, 30, 4 (August 2000), 345-354.
- [97] Shuster, L. I. and Ruscello, D. M. Evoking [r] Using Visual Feedback. *American Journal of Speech-Language Pathology*, 1, May 1992), 29-34.
- [98] Strik, H. and Cucchiari, C. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29, 2-4 1999), 225-246.
- [99] Studiocode Business Group. Studiocode business group - supplier of studiocode and stream video analysis and distribution software. <http://www.studiocodegroup.com> Accessed 2007
- [100] Suchman, L. A. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, New York, NY, 1987.
- [101] Tartaro, A. Storytelling with a virtual peer as an intervention for children with autism. *SIGACCESS Access. Comput.*, 84 2006), 42-44.
- [102] 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.
- [103] Viégas, F. B. and Donath, J. S. Chat circles. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* (Pittsburgh, Pennsylvania, United States, 1999). ACM, 1999.
- [104] Wetherby, A. M., Prizant, B. M. and Hutchinson, T. A. Communicative, Social/Affective, and Symbolic Profiles of Young Children With Autism and Pervasive Developmental Disorders. *American Journal of Speech-Language Pathology*, 7, May (May 1998), 79-91.
- [105] Whyte, W. *The social life of small urban spaces*. Conservation Foundation, Washington, DC, 1980.
- [106] 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.

Appendix A: A³ Coder's Guide

A³ CODER GUIDE

*denotes ranged event

PASS 1: Use Standard Playback Mode

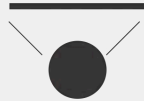
***Non-Child Audio:** Durations when audio/sound made & see on bars (you can hear it) and is NOT coming from the child /overlaps with child with child's sounds OR sound caught on volume bars, but is not identified as from child (unknown source). Include REGULAR heavy breathing. Mark whole segment if "contaminated"

PASS 2: Use Interval Payback Mode (Continuous Playback)

Smiling: When the child appears to be smiling during the past 3 seconds, and in the past 3 seconds, we at some point could see his face.

No Face: Can not see face (to determine smile) at all in the past 3 sec.

Oriented @ Screen: When the child, in the past 3 seconds, was facing towards the screen within 90 degrees. Spinning through the 90 degree arc should not be counted. Rather, time where the facing direction is within the arc for at least a "moment."



Auditory Focus: During the time frame, did the child get closer in proximity to, or touch the speaker. OR Child is not in the visual arc, in response to computer sounds, orient to the visual arc.

PASS 3: Use Standard Playback Mode

Use a 2 second pause between end and start to delineate between vocalizations
Also mark sounds even if not recorded by computer

***Laugh:** The sound should NOT be able to be transcribed as a speech like vocalization. To qualify as laughter it needs to be paired with a positive affect.

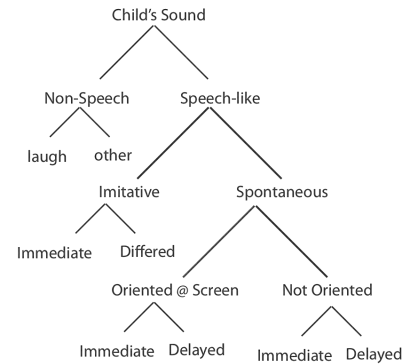
Non-Speech Vocalization: When a vocalization made is non speech, and is not a laugh. Includes gulps, screech, grunt, lip pops, ticks, heavy sighs, etc. Lasts for 2 seconds before code again.

***Speech Like Vocalizations:** When a vocalization is made by the child, the phonetic construction of the sound should be noted as an event. Marked at the start of the sound
Use the annotation hotkey to write the sound made
New Sounds are formed by gap of 2 seconds, OR separated by a non-speech sound, or laugh, or computer sound (while the child is not making a speech-like sound).

Turn taking: the computer makes a sound, and then the child starts sound if nucleus of final syllable (vowel) has been initiated
Speech Like Sounds Only

BIGmack Switch: When a child presses the switch, mark this at the start of the press. Do not mark when switch is pressed by anyone else.

For each speech like vocalization mark one of the following Correspondences.



Spontaneous: creating a non-imitative vocalization

I+S spontaneous (Immediate, Oriented @ Screen, Spontaneous): while oriented towards screen (within in 2 seconds of starting sound), speaker, within 5 seconds of end of source sound

D+S spontaneous (Delayed, Oriented @ Screen, Spontaneous): while oriented towards screen (within in 2 seconds of starting sound), speaker, after 5 seconds of end of source sound

I spontaneous (Immediate Spontaneous): while NOT oriented towards screen (within in 2 seconds of starting sound), speaker, within 5 seconds of end of source sound

D spontaneous (Delayed): while NOT oriented towards screen (within in 2 seconds of starting sound), speaker, after 5 seconds of end of source sound

Imitative: the child attempts to echo or repeat a sound previously heard/made by computer or human. Must match 50% of phonemes of ATTEMPTED target OR same number of syllables as whole. Target resets after each new non-child sound. Cannot imitate self OR echoed sound

Differed imitation: time frame for Differed Imitation ends after a new sound is made by speaker, or researcher

Immediate imitation: within 5 seconds from source

PASS 4: Drag Position for fast skim

***Time In Chair:** marking times during the video when the child is seated in the chair. This includes having the child's butt on the chair, 2 legs on the chair (ie. sitting cross-legged, sitting on feet) or otherwise in the chair in a manner generally deemed "sitting". Not included is one leg on chair, one leg standing.

Appendix B: Letter of Copyright Permission



Advancing Computing as a Science & Profession

Publications

Department of

October 14, 2008

Joshua Hailpern,
Ph.D. Candidate
Department of Computer Science
University of Illinois at Urbana-Champaign

Re: Permission to Reprint ACM Material in PhD Dissertation or Thesis

Dear Mr. Hailpern:

In reply to your recent request to reprint excerpts from the following ACM copyrighted material in your dissertation:

VCode and VData: Illustrating a New Framework for Supporting the Video Annotation Workflow,
Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2008) pp. 317-321.

<http://doi.acm.org/10.1145/1385569.1385622>

A3: A Coding Guideline for HCI+Autism Research Using Video Annotation. Proceedings of the 10th Intl
SIGACCESS Conference on Computers and Accessibility (ASSETS 2008), pp. 11-18.

<http://doi.acm.org/10.1145/1414471.1414476>

ACM is pleased to grant the permission without fee for your use as requested. This includes gratis permission for your university library to provide individual copies of this material for personal use and interlibrary loan as part of your dissertation.

However, subsequent use of the articles by UMI or other document delivery providers requires an appropriate license and payment of the per copy fee to the Copyright Clearance Center (www.copyright.com). Other republication or commercial use of the paper requires a separate permission and/or fee.

The original citation, ACM copyright notice and notice of permission are required to appear on the first page of the reprint in any media, and the Digital Object Identifier (DOI) should be included as a link to the definitive article in the ACM Digital Library.

Please don't hesitate to contact me at 212-626-0652 or cotton@hq.acm.org if you have further questions.

Sincerely,

Deborah Cotton

Deborah Cotton,
Copyrights and Permissions

2 Penn Plaza, Suite 701
New York, NY 10121-0701

Tel: +1-212-869-7440
Fax: +1-212-944-1318

acmhelp@acm.org
www.acm.org

Author's Biography

Joshua Hailpern was born in Katonah, New York in 1983. Joshua Hailpern is currently a Ph.D. student at the University of Illinois at Urbana-Champaign in the Computer Science Department focusing on Human Computer Interaction. He attended Carnegie Mellon University in Pittsburgh, PA, majoring in Computer Science and minoring in Music Technology and graduated with a Bachelor of Science degree with Honors in May 2006.