A³: A Coding Guideline for HCI+Autism Research using Video Annotation

Joshua Hailpern
Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61802 USA
1-217-333-3328
Jhailpe2@cs.uiuc.edu

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

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

Laura DeThorne
Speech & Hearing
University of Illinois
901 South Sixth St
Champaign, IL 61820 USA
1-217-333-2230
lauras@uiuc.edu

Mary-Kelsey Coletto
Speech & Hearing
University of Illinois
901 South Sixth St
Champaign, IL 61820 USA
1-217-333-2230
mcoletto@uiuc.edu

ABSTRACT

Due to the profile of strengths and weaknesses indicative of autism spectrum disorders (ASD), technology may play a key role in ameliorating communication difficulties with this population. This paper documents coding guidelines established through cross-disciplinary work focused on facilitating communication development in children with ASD using computerized feedback. The guidelines, referred to as A³ (pronounced A-Cubed) or **A**nnotation for **A**SD **A**nalysis, define and operationalize a set of dependent variables coded via video annotation. Inter-rater reliability data are also presented from a study currently in-progress, as well as related discussion to help guide future work in this area. The design of the A³ methodology is well-suited for the examination and evaluation of the behavior of low-functioning subjects with ASD who interact with technology.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation/ methodology

K4.2 [**Social Issues**]: Assistive technologies for persons with disabilities

General Terms

Measurement, Reliability, Experimentation, Human Factors

Keywords

Autism, ASD, Coding, Guideline, Video, Annotation, Reliability, Point-by-Point agreement, Kappa Audio Feedback, Visualization

1. INTRODUCTION

During a child's typical development, speech and communication skills appear to unfold effortlessly. For some populations, however, basic communication skills remain a lifelong struggle. One such demographic is children who are diagnosed as being "low-functioning" with Autism Spectrum Disorder (ASD). Communication, empathy, social function, and expression can all be impaired in individuals diagnosed with ASD. Many of these individuals will struggle their entire life with social interaction as well as basic communication skills. The CDC reports that 1 in 150 children will be diagnosed with ASD, and current trends show this number is increasing [7]. Without treatment, speech skills, as well as other forms of interpersonal interaction, may be substantially impaired [26], leaving individuals with limited means of communicating basic wants and needs. Critical research is being conducted, focusing on skill acquisition, and language formation [3, 11, 19, 26, 30, 32].

Technology is situated to increase the impact of intervention and research for those diagnosed with ASD. Work conducted to date has explored diagnosis [22], play [31, 34], audio perception [41], and interpersonal skills for high functioning children with ASD [46]. 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 paper, 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 **A**nnotation for **A**SD **A**nalysis to quantitatively assess a set of dependent variables identified through the digital video annotation process. Through the application of A³ in a research context, we demonstrate the inter-rater reliability of the annotations, as well as directions for its improvement. Because we are 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 work is in the demonstration of a new coding system, its reliability, and an overview of the results found when it was applied in the context of an actual study.

2. RELATED WORK

Analysis of human behavior can be traced back hundreds of years [8] 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 [24, 39, 40]. 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 [23, 33, 47, 49], interaction (with and without physical presence) [3, 47], diagnosis via observation [25] and communication skill acquisition [35, 44]. 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 [17, 28, 43].

Computer Scientists, particularly in HCI, have developed a broad set of coding guidelines for a large array of tasks [45, 48]. 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 [6, 46] and most, gathered data through subjective (qualitative) observation [21, 31, 34].

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 paper addresses this gap by detailing the construction of A^3 and the reliability of the variables in an experimental research setting.

3. EXPERIMENTAL SETTING

A four-month study began in the Fall of 2007, examining the effect of different forms of audio and visual feedback on a multiplicity of responses (motor & verbal behavior) by low-functioning children with autism [14]. By reinterpreting the subject's vocalizations into visual and auditory feedback, the study examined the child's behavioral response to these forms of feedback. A simple example is a circle that changes in diameter as volume changes, or a clip of music/sound (whose duration is proportional to the length of sound produced) that is played back to subjects when their sound production ceases. Five "low functioning" children (aged 3-8 years) were enrolled in this pilot study for six, 40-50 minute videotaped sessions, each approximately one week apart. Each session consisted of eight two-minute trials, on average, in which the subject was presented with a permutation of audio and/or visual feedback contingent on sounds produced by the subject. A control condition (neither audio nor visual feedback) was conducted at the beginning of each session to serve as a comparison between or an anchoring point for the data collected.

Two coders reviewed up to two session tapes per week (80-100 minutes of data). From each session, one 2-minute trial was annotated by both coders, and was discussed in a weekly reliability meeting. Discrepancies were examined together by researchers and coders. The result and implications of these agreement values are discussed later in this paper.

4. A^3: REFINEMENT METHODOLOGY

The study 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.

4.1 VCode and VData

As we defined and refined our initial set of coding guidelines, we examined extant video annotation tools. Upon a thorough exploration of the tool set available (both free and commercial), none satisfied our criteria; specifically, existing tools appeared to be overly complicated, possessing too many features that were not critical for annotation (e.g., transcription), did not support agreement calculations, or utilized poor interaction techniques. In addition, there were sets of features necessary for annotation that were non-existent or poorly supported [13]. Most critically, however, existing tools did not facilitate the coding workflow: **1)** establish video clips and guidelines; **2)** training of coders and easy real-time reliability checks; **3)** annotation of videos; **4)** weekly reliability checks; **5)** repeat steps 3 and 4, ad infinitum; **6)** exporting of data to statistical packages.

To resolve the discrepancy between researchers' needs and existing tools, two tools were created, *VCode* (for annotation of videos) and *VData* (for agreement and reliability checking), which were explicitly designed to facilitate the coding practices of researchers in Computer Science and the Behavioral Sciences (e.g. Speech and Hearing Science and Special Education) [13]. *VCode* and *VData* (Figure 1) were released as an open-source free download online for public use (http://social.cs.uiuc.edu/projects/vcode.html).

Figure 1: *VCode* screen-shot (center), *VCode* Admin Window (left), *VData* (right), logo (bottom)

In addition to supporting the coding workflow, enabling rapid agreement checks, linking annotations to video, and standardizing of material annotated by coders, *VCode* facilitated multiple modes of playback, including traditional or Continuous Playback and Interval Playback. Continuous Interval Playback Mode is a feature of the *VCode* analysis system that plays a video for X seconds, then pauses, allowing the coder to ask if the specified event occurred anywhere in the past X seconds. This is ideal for variables that are not discrete (i.e., difficult to specify their exact starting or ending point, or exact duration). When checking agreement values, *VData* allows researchers to dynamically adjust the agreement tolerance, thus accommodating for mark placement discrepancies. Because of this enhanced feature set, these tools were utilized in the refinement of the coding guidelines, annotation process, and agreement checks. Though *VCode* and *VData* were essential means to annotate video, test, and validate the reliability of A^3, their development is not the focus of this work.

4.2 Process of Refinement

Although the initial set of coding guidelines was created during the experimental process, the final set presented in Appendix 1 and Section 5 were created through a 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 [20], 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 [20]. 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. A^3: DESCRIPTION

The following section is a detailed description of the dependent variables examined in the A^3 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 1. 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.

5.1 Motor Behavior Variables

With the exception of the metric *Time In Chair* (5.1.5), which was gathered with standard playback, all these metrics were gathered with the Continuous Interval Playback Mode set to three seconds.

5.1.1 Smiling

The variable, *Smiling*, was chosen because it is typically associated with pleasure or enjoyment (e.g. [11]). 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.

5.1.2 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).

5.1.3 Oriented at Screen

In order to assess visual attention to content, we created an "orientation arc" for the evaluation of child gaze; see [3, 11] 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 (~90°) was used to accommodate the behavior in which children with autism will use their peripheral vision as primary visual input [18]. See Appendix 1 for illustration of the orientation arc.

5.1.4 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 [9, 29].

5.1.5 Time in Chair

To asses the willingness to attend to computer stimuli, we coded the duration a child would spend in his/her chair [27, 42]. We hypothesized that increased time sitting was a proxy for engagement.

5.2 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 1). With the exception of *Turn*

Taking (5.2.7), 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.

5.2.1 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 which 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) [49].

5.2.2 Non-Speech Sounds (Laugher 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 [12]. 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 [44]. By annotating non-speech sounds, we also hoped to examine the impact of the external stimuli on their non-speech vocalization.

5.2.3 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)[16]. 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.

5.2.4 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 [35]. This distinction is particularly relevant for children with autism due to echolalic tendencies [36]. Theory suggests that words/sounds in delayed imitation are stored outside of the subjects short-term or working memory.

5.2.5 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 [3, 47]). 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.

5.2.6 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).

5.2.7 Turn Taking

An important skill in oral communication is that of *turn taking*, or waiting for others to finish [33, 47]. 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 Other Metrics

There were two other metrics collected not expressly discussed in section 5.1 and 5.2; BIGmack Switch and *Non-Child Audio*.

5.3.1 BIGmack Switch

The BIGmack™ Switch [1] is an assistive technology device used to play a pre-recorded message for individuals with speech disability [37]. 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. However, at the time of publication, we had not completed data analysis for this particular subject. As a result, no data are reported here.

5.3.2 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.

6. RESULTS

Overall, agreement between coders was acceptable (above 80%) for most of the variables. By the last video, coders spent approximately 20 minutes per one minute of video footage annotating, a significant decrease 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 217 trials from four of the subjects (at the time of this paper submission, data collection and analysis for the fifth subject had not been completed). 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 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 [20] 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 [38]. The first Kappa accounts for the level of observer agreement on whether they

Table 1. Percent Agreement for Dependent Variables

Variable	% Agreement
Non child audio	76.71
Duration: Non child audio	85.00
Smiling	91.20
No face	81.00
Oriented at screen	98.45
Auditory focus	85.53
Laugh	51.85
Duration: Laugh	85.71
Non speech vocalization	84.53
Speech like vocalization	89.95
Duration: Speech like vocalization	94.71
Turn taking	82.00
Immediate + Screen + Spontaneous	78.76
Delayed + Screen + Spontaneous	78.26
Immediate + Spontaneous	68.18
Delayed + Spontaneous	71.43
Differed imitation	*None Recorded*
Immediate imitation	83.33
Time in chair	86.84
Duration: Time in chair	78.79

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.

6.1 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 eight of the 20 measured variables had IRA that exceeded 85%, 12 had IRA that exceeded 80%, and 16 exceeding 75%. A further exploration of variables with less than 80% agreement is provided in the following section (7.1).

The Kappa statistics calculated from the data suggest a high level of agreement. Kappas ranged from 0.69 (Good) to 0.81 (Very Good). Our interpretation of agreement follows from that of Altman and Byrt [2, 5].

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.

Table 2. Kappa statistic for variables coded using interval playback of 3 seconds.

Variable	Kappa
Auditory focus	0.63
Oriented at screen	0.81
No Face	0.73
Smiles vs. No smiles	0.69

7. 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 A^3 system on the outcomes of the original study for which it was the system of measurement.

7.1 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 A^3 system.

7.1.1 Laughter

The Laughter variable had one of the lowest agreement values (51.85%) 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 [4].

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 [47].

7.1.2 Spontaneous Speech-Like Vocalizations

The sub-divisions within *Spontaneous Speech-Like Vocalizations* also resulted in low agreement between coders (68.18% to 78.76%). 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

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

Variable	% Agreement
Immediate spontaneous	80.00
Delayed spontaneous	81.11
All spontaneous	82.78

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."

7.1.3 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.

7.2 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.

7.3 Child Differences

Overall, data gathered with our interactive feedback displays have been used to reveal meaningful results in the initial analysis. Specifically, the feedback facilitated speech-like vocalizations in one child compared to a no feedback condition, with a similar trend noted in a second child that did not reach statistical significance [10]. Post-hoc analyses on three subjects suggested that auditory feedback was more effective than visual feedback or mixed feedback in facilitating vocal use. Similar analyses are being conducted for the other participants with the added focus on measures of engagement. Other preferences towards specific feedback features (e.g. movement patterns of onscreen objects) have emerged as well [15]. These data, in conjunction with the findings of this study, will lead directly to future work in targeting specific types of sound production.

8. CONCLUSION

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 A^3 (**A**nnotation for **A**SD **A**nalysis). 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 paper demonstrated A^3'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. However, not all variables were coded with equal success. For future experiments, we hope to capitalize on the lessons learned and improve the descriptions of some variables (such as *Smiling* and *Laughing*), while also improving audio recording techniques.

9.ACKNOWLEDGMENTS

We would like to thank all of the participants and their families, our four terrific coders (Ashley Sharer, Christine Holloway, Sammi Goldenberg, Christine Renee Birn), NSF (NSF-0643502), our friends, family, and loved ones.

10.REFERENCES

[1] AbleNet® BIGmack® communicator - Red. http://www.ablenetinc.com/item_detail.aspx?ItemCode=10000201 Roseville, MN, 2008

[2] Altman, D. Practical Statistics/or Medical Research. Chapman and Hall, London, 1991.

[3] 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.

[4] 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.

[5] Byrt, T. How good is that agreement? Epidemiology, 7, 5 (September 1996), 561.

[6] 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.

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

[8] Clifford, J., Marcus, G. E. and School of American Research Writing Culture. University of California Press, Berkeley, California, 1986.

[9] 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.

[10] DeThorne, L. S. and Coletto, M. Visualizing voice: Use of computerized feedback to facilitate vocalizations in children with autism. Proseminar in Speech and Hearing Science University of Illinois at Urbana-Champaign. Champaign, IL, April 2008.

[11] 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.
[12] 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.
[13] Hagedorn, J., Hailpern, J. and Karahalios, K. G. VCode and VData: Illustrating a new Framework for Supporting the Video Annotation Workflow. In Proceedings of the Advanced Visual Interfaces (Napoli, Italy, 2008). ACM-PRESS, New York, NY, 2008.
[14] Hailpern, J. Encouraging Speech and Vocalization in Children with Autistic Spectrum Disorder. SIGACCESS NEWSLETTER (Sept 2007, 89) 47-52.
[15] Hailpern, J., Karahalios, K., Halle, J., DeThorne, L. and Coletto, M. Creating a Spoken Impact: Audio and Visual Feedback for Children with Autistic Spectrum Disorder. In Proceedings of ([In Preparation]), [In Preparation].
[16] 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.
[17] 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.
[18] Howlin, P. An Overview of Social Behavior in Autism. Plenum, New York, NY, 1986.
[19] Kanner, L. Autistic Disturbances of Affective Contact. V.H. Winston, 1943.
[20] Kazdin, A. E. Single-Case Research Designs: Methods for Clinical and Applied Setting. Oxford University Press, USA, 1982.
[21] Kerr, S. J., Neale, H. R. and Cobb, S. V. G. Virtual environments for social skills training: the importance of scaffolding in practice. In Proceedings of the Proceedings of the fifth international ACM conference on Assistive technologies (Edinburgh, Scotland, 2002, 2002). ACM Press, New York, NY, 2002.
[22] Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N. and Abowd, G. D. Grow and know: understanding record-keeping needs for tracking the development of young children. In Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems (San Jose, California, USA, 2007). ACM Press, New York, NY, 2007.
[23] 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.
[24] Lee, L. Developmental Sentence Analysis. Northwestern University Press, Evanston, IL, 1974.
[25] 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.
[26] Lovaas, I. I. The Autistic Child. John Wiley & Sons, Inc, New York, 1977.
[27] Lovaas, O. I. Teaching Individuals with Developmental Delays: Basic Intervention Techniques. PRO-ED, Inc., Austin, TX, 2003.
[28] Luo, Y. and Baillargeon, R. Can a Self-Propelled Box Have a Goal? Psychological Science, 16, 8 (2005 2005), 601-608.
[29] 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.
[30] McGee, G. G., Krantz, P. J. and McClannahan, L. E. The facilitative effects of incidental teaching on preposition use by autistic children. Journal of Applied Behavior Analysis, 18, 1 (Sprirng 1985), 17-31.
[31] Michaud, F. and Théberge-Turmel, C. Mobile robotic toys and autism. Springer, 2002.
[32] National Research Council. Educating Children with Autism. Division of Behavioral and Social Sciences and Education, Washington, DC: National Academy Press, 2001.
[33] Owens, R. E. Language Development: An Introduction (7th Edition) Allyn & Bacon, Boston, MA, 2007.
[34] 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.
[35] Prizant, B. M., Schuler, A. L., Wetherby, A. M. and Rydell, P. Enhancing language and communication: Language approaches. Wiley, New York, 1997.
[36] Rapin, I. and Dunn, M. Language disorders in children with autism. Seminars in Pediatric Neurology, 4, 2 (June 1997), 86-92
[37] Reichle, J., Beukelman, D. and Light, J. Implementing an augmentative communication system: Exemplary strategies for beginning communicators. Brookes Publishing Company, Baltimore, MD, 2002.
[38] 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.
[39] Retherford, K. S. Guide to Analysis of Language Transcripts. Thinking Publications, Eau Claire, Wisconsin, 1993.
[40] Rosenblum, K. L., 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 (December 2004), 544-549.
[41] Russo, N., Larson, C. and Kraus, N. Audio–vocal system regulation in children with autism spectrum disorders. Experimental Brain Research, [Epub ahead of print - 2008] ([Epub ahead of print - 2008] [Epub ahead of print - 2008]).
[42] 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.
[43] 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.
[44] 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.
[45] Suchman, L. A. Plans and situated actions: The problem of human-machine communication. Cambridge University Press, New York, NY, 1987.
[46] 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.
[47] 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.
[48] Whyte, W. The social life of small urban spaces. Conservation Foundation, Washington, DC, 1980.
[49] 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 1. A^3 Coding Guidlines

A^3 CODER GUIDE

*denotes ranged event

PASS 1: *Use Standard Playback Mode*

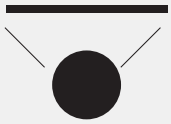
*Non-Child Audio: Durations when audio/sound made & see on bars (you can hear it) and is NOT coming from the child /overlaps with child with child's sounds OR sound caught on volume bars, but is not identified as from child (unknown source). Include REGULAR heavy breathing. Mark whole segment if "contaminated"

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.

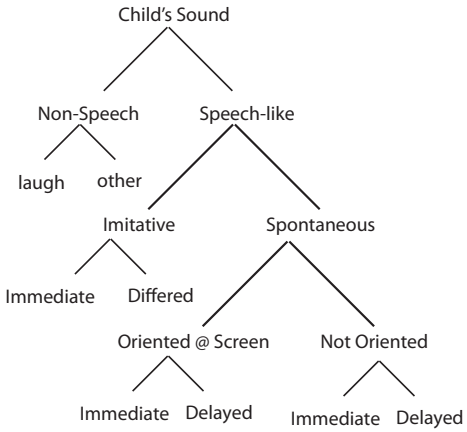
Non-Speech Vocalization: When a vocalization made is non speech, and is not a laugh. Includes gulps, screech, grunt, lip pops, ticks, heavy sighs, etc. Lasts for 2 seconds before code again.

*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.