Universal Access in the Information Society
https://doi.org/10.1007/s10209-019-00707-5

LONG PAPER

Improving communication skills of children with autism through support of applied behavioral analysis treatments using multimedia computing: a survey

Corey D. C. Heath[1] · Troy McDaniel[1] · Hemanth Venkateswara[1] · Sethuraman Panchanathan[1]

Abstract
Naturalistic applied behavior analysis (ABA) techniques have been shown to help children with autism improve their communication skills. Recognizing that individuals who interact with children regularly are in the position to utilize treatments with profound effects, researchers have examined methodologies for training parents, teachers, and peers to implement treatments. These programs are time intensive and often unable to support trainees after training. Technologies need to be examined to determine how they can aid in the educational and support process. Academic publications and publicly available training programs were reviewed to determine the types of participants, methodologies, and training durations that have been reported for instructing interventionists. These resources illustrate a need to make programs more accessible. To address this, selected computer science research is applied to methods of evaluating ABA implementations in order to recommend how the technologies could be utilized to make training and support programs more accessible. Review results of instructional programs, both in research and available in the community, illustrate the challenges in providing training in ABA methodologies. Modern research in multimedia data processing and machine learning could be applied to reduce the human cost of training and support individuals implementing ABA techniques. Utilizing machine learning techniques to analyze video probes of naturalistic ABA treatment implementation could alleviate the human cost of evaluating fidelity, allowing for greater support for individuals interested in the treatments. These technologies could be used in the future to expand data collection to provide more perspective on the treatments.

Keywords Applied behavior analysis · Pivotal response treatment · Autism spectrum disorder · Parent training · Multimedia data processing · Machine learning

1 Introduction

In early childhood development, people with the most regular interaction with the child have the most profound effects. Children learn the essential skills for life from their parents, relatives, and other individuals they interact with on a daily basis. Typically, these educational activities are intrinsically grounded in social interactions, observations, and consistent routines. Children with developmental disabilities, such as autism spectrum disorder (ASD), may have difficulty grasping the pivotal concepts from these social interactions. Applied behavior analysis (ABA) techniques have been developed to provide training for children with developmental disabilities.

Applied behavior analysis is an approach to creating and implementing procedures to promote beneficial behaviors or diminish disadvantageous behaviors. ABA focuses

✉ Corey D. C. Heath
corey.heath@asu.edu

Troy McDaniel
troy.mcdaniel@asu.edu

Hemanth Venkateswara
hemanthv@asu.edu

Sethuraman Panchanathan
panch@asu.edu

[1] School of Computing, Informatics and Decision Systems Engineering, Arizona State University, 699 S Mill Ave, Tempe, AZ 85251, USA

Published online: 08 January 2020

on applying a scientific approach to behavior treatments, emphasizing collection and analysis of data. Despite the emphasis on empirical approaches, ABA remains focused on the individual subject of the treatment, not on a research agenda. Implementation of ABA requires the interventionist to analyze and adapt the program to ensure that the individual subject achieves the greatest benefit. This adaptability helps facilitate one of the important aspects of ABA treatments—the ability to be generalized to target different behaviors under differing circumstances [2].

There are two general approaches to designing ABA treatments—contrived and naturalistic [45]. Contrived techniques, such as discrete trial teaching (DTT), involve controlled, structured learning activities selected by the individual administering the training. Naturalistic techniques rely on following the recipient's interests and incorporating learning objectives into the activity. This involves using the activity as a motivator for compliance. Contrived methods often use motivators that are not directly associated with the learning activity.

Much of the research into training caregivers, teachers, peers, and paraprofessionals in ABA has focused on naturalistic methods such as pivotal response treatment (PRT) and Early Start Denver Model (ESDM). These studies have shown that implementation of naturalistic ABA methods not only helps children improve social and communication skills, but they also help promote positive affect for both the child receiving the treatment and the adult providing it. Because of this, it is important to create systems that train and support individuals involved in the lives of children with ASD in naturalistic ABA treatments. The costs associated with training make it difficult to provide to all the individuals that need it. This will likely be exasperated as diagnoses of ASD continue to increase. It is therefore important to look at the ways technology can be incorporated into the system to ensure people have access to training materials and implementation feedback.

The intent of this paper is to examine naturalistic ABA interventions and training strategies to identify areas that could be supported by current technology. In particular, the focus is on technologies that help evaluate communication opportunities provided by interventionists. To accomplish this, the following research questions were explored:

- What are the current approaches to training non-clinicians in naturalistic ABA methodologies?
- What are the current evaluation strategies for assessing fidelity to implementation for individuals learning naturalistic ABA methodologies?
- What are the potential barriers potential trainees encounter that prevent access to training and support resources?
- What are the costs for clinicians that restrict the amount of training and support they can provide?
- What are the current advances in computer science that could alleviate costs and barriers restricting training and support resources?
- How can these technologies be applied to create an automated data analysis and feedback system for non-clinician implemented naturalistic ABA?

The following section will present the important components of naturalistic ABA techniques and supporting research, with much of the research focusing on PRT implementation for improving social and communication skills. Following the discussion on implementation, research regarding training non-clinicians, caregivers, teachers, peers, and paraprofessionals, for example, will be examined, including studies incorporating technology. The concluding sections of the paper will present new directions that should be explored in regard to current technology for supporting individuals implementing naturalistic ABA.

2 Methods

The research presented in the subsequent sections of this paper was obtained through queries on the Google Scholar and OneSearch databases. The citations utilized for explaining the computer science concepts are meant to provide examples of how technologies are implemented, their capabilities, and their limitations. Due to the breadth of available research, this is not a comprehensive review of these topics; rather, these citations are provided to show how these concepts could be applied. Information regarding communal programs was obtained from the web and was accurate based on the time of access in 2018.

3 Naturalistic applied behavior analysis implementation

Naturalistic ABA methods focus on keeping the recipient of the treatment active and making the learning activities relevant [104]. Implementation is based on a generalized three-part sequence of antecedent, response, and consequence. The antecedent focuses on the actions the treatment interventionist takes to prompt the child with a learning activity. First, this involves gaining the child's attention, which generally involves seizing control of the toy or activity the child is currently participating in. After gaining the child's attention, the interventionist can then give an instruction. Verbal instructions can include modeling the word or phrase the child is expected to say, saying the beginning of a sequence, such as counting, then expecting the child to say the final word, or providing a choice. A time-delayed prompt can also be used where the interventionist seizes the motivator and

waits for a response that has been previously modeled [57, 59]. Verbal instructions should be limited to the speaking level of the child.

The response is the child's reaction to the antecedent. Ideally, the child will respond by making an attempt at speaking the intended word or phrase. All genuine attempts are treated as correct. How complete the response should be is dependent on the child's current skill level. If the child is mostly non-verbal, a correct response could be an attempt at speech or vocalizing the phoneme of the expected word. If the child has previously demonstrated they can speak the word or phrase being prompted, the response should be at that level.

Consequence is the reward for complying with the instruction. Generally, this reward is the continuation of the activity the child was engaged in prior to the instruction. The interventionist should provide the reward as quickly as possible following an acceptable attempt at the learning objective to prompt compliance. The interventionist should also be contingent on the child providing an adequate attempt for his or her current skill level.

Outside of the antecedent, response, and consequence sequence, children should also be rewarded for initiating social interactions, asking questions, and spontaneous speech [65, 104]. Children with ASD can exhibit a deficit in initiating social interaction, and part of the naturalistic intervention should include creating situations that necessitate the child taking the initiative. This can include placing a favorite toy in a visible but unreachable location to encourage them to ask for it.

Learning objectives can be sorted into two types—target skills and maintenance skills. Target skills are new objectives the interventionist is presenting to the child in order to increase his or her ability. Skills that have been achieved by the child can become maintenance skills. Maintenance skills are intermixed with target skills during treatment sessions to ensure that the child continues to practice and to keep the child motivated by giving them a challenge they can easily overcome.

There are nearly 30 years of published research on naturalistic ABA, primarily PRT, focusing primarily on children with autism between the ages of 6–11 [136]. These studies have shown that by implementing PRT, children with autism demonstrate improvement in vocal communication and spontaneous speech that was generalized to scenarios outside the training context. In addition to language outcomes, studies examined how PRT affects the stress, motivation, and happiness of both parents and children.

Outcomes of studies examining language, social, and play skills honed in children through naturalistic ABA have been favorable. Improvement based on language assessments and social interactions was shown after PRT interventions by [51–55, 64, 65, 108, 129]. Improvement in joint attention after naturalistic ABA intervention was reported by [130]. Increases in social and symbolic play were published by [115, 116, 124]. Studies examining the affective state of children also showed positive results following treatments. PRT was correlated with a reduction of anxiety in children by [71], resulting in a reduction of disruptive behaviors.

Studies conducted by [58, 62, 82] compared PRT to DTT or a similar contrived ABA implementation to evaluate children's post-intervention communication skills. Looking at mean number of spontaneous utterances, intelligibility, and mean length of vocal utterance, respectively, each study concluded that children that received PRT showed greater improvement than children who received a contrived ABA implementation. Similar conclusions were drawn regarding reduction of disruptive behaviors by [60, 83] with children in the PRT treatment group showing a greater reduction over adult-led interventions. In addition to language and behavior, affect was examined in two studies [61, 103] which concluded that PRT was related to greater increases in happiness and reduction of stress of both parents and children compared to DTT interventions.

4 Implementation of interventions by non-clinicians

If left only to clinicians to implement, the impact for treatments would be reliant on the amount of time the clinician could spend with the subject. To make naturalistic ABA more impactful, it is important to train caregivers, teachers, peers, and paraprofessionals that interact with the subjects more frequently in intervention methodologies. Research revealed that not only can non-clinical professionals learn to implement naturalistic treatments that improve child outcomes, but also that participating in these outcomes leads to improved affect for both the interventionist and the child.

Studies focusing on training parents of children with ASD to implement interventions for improving the child's communication skills illustrated that parents could effectively learn the techniques and display a high degree of implementation fidelity. The child's improvement on language assessment was often correlated with the implementation fidelity of the parent [4, 15, 29, 35, 68, 112]. The improvements associated with PRT training for parents was concluded to be independent of age, gender, or ethnicity [4]. Attempts to start interventions early in the child's development have also fueled research into training parents to implement interventions. Positive effects for infants after parents received naturalistic ABA training were reported by [56, 117].

In addition to communication skills, caregiver-implemented interventions have been studied for improving joint attention and have been the focus of research publications. Joint Attention, Symbolic Play, Engagement, and Regulation

(JASPER) is an intervention technique that seeks to utilize a child's interest in a toy or activity to practice socialization, verbal and gesture communication, and play behaviors. Parent-implemented JASPER interventions have been shown to improve joint attention skills in preschool-age children with ASD [44, 46]. Teacher implementation of JASPER, also on preschool-age children with ASD, also showed positive effects on joint attention, with noteworthy effects on child-initiated joint attention [69].

Beneficial effects on parents were concluded from the studies in addition to improvements in the children's language and social interaction skills. A reduction of stress levels and an increase in satisfaction were noted after training in PRT [117] and ESDM [24]. This is particularly important since parents of children with ASD report high levels of stress [43, 46, 85]. This high level of stress can also affect the behavior of the child in addition to the caregiver's well-being. Adding stress management has been shown to aid child outcomes and improve participation in treatments [46].

Peer implementation of PRT for elementary school students has also been explored [36, 89, 90]. Research studies indicate that peer interventions had a positive effect on social interaction and key behaviors that lead to making friends. It is also suggested that providing multiple peer interventionists could be beneficial for helping the treatment recipient generalize social skills [90]. Additionally, having multiple students working together to support their peer with ASD likely encourages students to become peer mentors and creates an enjoyable environment for the interactions [36].

5 Training non-clinicians in ABA

Parsing out training procedures and time spent on training from the presented research is difficult due to non-standardized reporting techniques, different baseline knowledge from the parents, utilization of different materials and methods, and individualized training durations for participants in the same study. Additionally, it is presumed in most cases that the interventionist-in-training was provided feedback after sessions recorded for data collection. Table 1 shows the training and intervention duration from 20 studies; however, these times do not include self-study times when the trainee was provided written materials. The average duration was 7.6 h, with the most common duration being 12 h. The study from [44] was not included in these calculations. The participants in the study received 1 h of training prior to providing interventions; however, the participants are noted as receiving extensive training from their child's preschool prior to the study. Participants conducted a substantial number of intervention sessions, ranging from 54 to 290 sessions. The duration of these sessions was not reported.

The duration metrics from Table 1 illustrate the majority of the research studies require a significant commitment from both the participating trainees and the trainers. This time does not reflect the time that would be needed to analyze performance metrics that would be required to give pointed feedback. The training time recorded in these studies is lower than courses offered at community resource centers. A brief search of caregiver training programs from autism resource centers in the USA shows training options are typically centered around group workshops or one-on-one support sessions. For example, an 8-h group course teaching ABA is offered by the University of Washington Autism Center [127] in Seattle. Twenty-hour group courses are offered by The Help Group [123] in Sherman Oaks, California, and the Southwest Autism Resource and Research Center (SARRC) [114] in Phoenix, Arizona, on skills for parenting children with developmental disabilities and PRT implementation, respectively. The Children's Hospital at Sacred Heart [13] offers a 12-h ABA implementation course in Pensacola, Florida.

One-on-one training courses were advertised by SARRC and the Choice Autism Center in Traverse City, Michigan [14]. The SARRC Web site listed an individualized 12-session, 1-h-per-week course on ABA implementation along with an intensive one week course. The Choice Autism Center lists two individualized training programs based on the age of the child. For a child aged 18 months to 5 years, a 20–40-h-per-week program is listed. A 6–20-h-per-week program is available for children ages 6 to 12. Many other autism centers offered individualized in-clinic or in-home programs, but did not list specific durations.

Both the research studies and the available community programs indicate that a time commitment is expected when learning and performing ABA interventions. This could be problematic and may restrict accessibility for many people who need to learn the procedures. Many of the programs that were presented in community centers were intensive, requiring several hours per day. For working parents, this would mean taking time from work along with finding childcare.

The courses can be problematic for behavioral analyst instructors as well as trainees. In group settings, the instructor may not have the opportunity to provide sufficient feedback to each individual participant, either due to time constraints or privacy. One-on-one courses require the analyst to focus on one parent–child dyad for an extended period of time. While this is beneficial to the participants in the course, it is a difficult model to maintain due to the rising number of individuals needing training and assistance. Additionally, analysts often need to compile reports and feedback to provide to participants, adding additional time requirements on top of providing instruction.

The location of autism resource centers could also potentially limit participation. Many of the centers are associated

Table 1 Publications on training non-clinicians in ABA implementation

Publication	Training method	Participants	Training duration (hours)
Laski et al. [68]	In-person clinician instruction	Parents	3.75[a]
Pierce and Schreibman [89]	In-person clinician instruction	Peers	2
Pierce and Schreibman [90]	In-person clinician instruction	Peers	2
Koegel et al. [63]	In-person clinician instruction	Parents	25
Symon [120]	In-person peer instruction	Parents	25
Baker-Ericzén et al. [4]	In-person clinician instruction	Parents	12
Gillett and LeBlanc [29]	In-person clinician instruction	Parents	3
Harper et al. [36]	In-person clinician instruction	Peers	1
Jones and Feeley [44]	In-person clinician instruction	Parents	1[b]
Vismara et al. [131]	Teleconference/video instruction	Therapists	17
Coolican et al. [15]	In-person clinician instruction	Parents	6
Machalicek et al. [77]	Teleconference	Teachers	1.25[c]
Nefdt et al. [85]	Video instruction	Parents	1.6
Lawton and Kasari [69]	In-person clinician instruction	Teachers	6
Vismara et al. [132]	Teleconference	Parents	12
Vismara et al. [133]	Teleconference/video instruction	Parents	12
Steiner et al. [117]	In-person clinician instruction	Parents	10
Estes et al. [24]	In-person clinician instruction	Parents	12
Koegel et al. [56]	In-person clinician instruction	Parents	7.33[d]
Kasari et al. [46]	In-person clinician instruction	Parents	10
Gengoux et al., Hardan et al. [27, 35]	In-person clinician instruction	Parents	16
Smith et al. [112]	In-person clinician instruction	Parents	8
Suhrheinrich and Chan [118]	In-person clinician instruction/ video instruction	Teachers/para-professionals	18

[a]Maximum interventions were nine 25-min sessions

[b]Parents had prior PRT training

[c]Based on an average of 60-to-90-min sessions

[d]Average based on three participants that had 4, 6, and 12 h of training, respectively

with medical centers, universities, or research institutions, often located in larger metropolitan areas. Individuals in rural or remote locations may have not have a resource center in the immediate area and may not be able to travel to receive training.

6 Alternatives to in-person training

Research publications have sought ways to mitigate the limitations of in-person training by examining alternative means to training. A study conducted by [120] explored having caregivers who received PRT training teach their immediate co-caregivers intervention techniques. They found that the trainees were able to adequately learn and successfully implement PRT. While this is an interesting study on disseminating information, most of the research involving alternatives looked into the application of technology to facilitate distance learning. This was accomplished by the creation of digital self-directed learning platforms and live telecommunication broadcasting.

Vismara et al. [131] examined technology for remote instructions for training therapists to teach parents to implement ESDM. Their study organized the participating therapists into two groups with one group receiving live in-person training and the other group receiving training via remote video-conference sessions. They found that both groups performed equally well at instructing parents. Vismara et al. [132] applied this telecast training model to teach parents the ESDM methodology directly. The training consisted of live broadcast training sessions between clinicians and parents using video conferencing software. They found that parents were able to learn the techniques through the video conferences. They also showed that improvements in the child's engagement scores correlated with improvements in the parent's fidelity to implementation. Vismara et al. [133] combined the use of video-conference training with self-directed online resources. They found that online resources directly related to learning

more about ESDM were utilized more than other features, such as media sharing or calendar functions.

Video-conference systems for providing real-time feedback for teachers implementing ABA-based treatments in a classroom environment were explored by Machalicek et al. [77]. The teachers would set up the video equipment to broadcast a feed of the classroom to a remote expert who would provide instructional feedback during the session. They found that difficulties with setting up the required equipment impacted the success of the study. The technology was also distracting the students in the classroom and, at times, student behavior obstructed the communication between the teacher and the clinician. They concluded that the utility of this approach was largely dependent on the teacher's ability to set up and troubleshoot the equipment. They did not address issues with technology being distracting, initial reductions in fidelity after the baseline, and possible limitations to real-time feedback.

Video modeling of behavioral treatments in various scenarios was evaluated for training parents to implement DDT procedures [3]. These videos were designed to train the parents to implement DDT procedures with their children. The researchers' primary focus was on whether or not parents would comply with a training schedule consisting of video modeling. They concluded that 70.6% of the participants attended between 50 and 100% of the video sessions; however, no results are stated regarding how effective the video training was for improving target skills for parents or their children.

Also using a self-directed video platform, Nefdt et al. [85] explored training parents to implement PRT. Their results showed that the majority of participants completed the training and were able to demonstrate sufficient fidelity in implementing PRT in post-training evaluations. This result corresponded with an increase in child vocalizations. Additionally, the researchers reported that the caregivers showed greater confidence during the post-training intervention session.

Programmable robots were implemented in a study to explore their use as a means of fostering engagement in behavior treatment sessions for children with autism [28]. The robot was programmed with scenarios that were based on ABA implementation. The clinician could then select the pre-programmed scenarios the robot enacted based on the child's needs or preferences. The researchers concluded that the robot would need to be easily customizable and expandable in order to be a functional tool for implementing PRT training. The need for continuous adaptation and the concept of in-context learning made covering all the scenarios difficult. This underlines the difficulty of a fully autonomous system for conducting behavioral training.

7 Data measurements and fidelity to implementation

Regardless of whether the training is occurring in person or at a distance, the most common method of scoring fidelity of intervention implementation and providing feedback is the use of video probes. Typically, these video probes consist of 10-to-15-minute videos of the interventionist working with the child receiving the treatment. The overall time period is then broken into one-to-two-minute increments to be scored on fidelity. An intervention is considered to be performing aptly if they score over 80% [63]. The expectation is that interventionists are providing approximately two learning opportunities per minute.

Assessments of implementation fidelity are based on the three-part sequence of antecedent, response, and consequence described earlier. Although these categories are often adapted to fit the intervention methodology and the child skill being targeted, they typically consist of the following: creating an opportunity to respond, task variation, following the recipient, identifying natural motivators, contingency, and reinforcing attempts [63, 85].

Creating an opportunity to respond requires two key features. First, the interventionist must have the recipient's attention. Generally, this means that the recipient is not engaging in a solitary activity and is not exhibiting disruptive or self-stimulating behavior. Signs that the recipient is paying attention to the interventionist include looking at or in the direction of the interventionist, looking at an object being used for a shared activity, or reaching for an object in the interventionist's control [57, 70, 119]. Methods for gaining the recipient's attention should be focused on the interventionist incorporating himself or herself into the activity the recipient is engaged in. This allows the interventionist to have shared control of the activity to integrate learning opportunities. Calling the recipient's name or using physical contact to gain his or her attention should be kept to a minimum.

The second feature of creating an opportunity to respond is providing a clear instruction. This can take the form of either a verbal instruction or a gestural prompt, depending on the target skill and the abilities of the recipient. For communication skills, typical instructions are categorized as being a model prompt, a choice, a question, a lead-in statement, or a time delay. For model prompts, the interventionist speaks the word the recipient should attempt. Choice instructions include giving the recipient two or more options based on the motivator with the intention that he or she makes a vocal attempt at one of them. Question instructions prompt the recipient to formulate a response based on the context. Lead-in statements present a known sequence, such as "ready, set, go," with the final

word, in this case, "go" being omitted by the interventionist with the intention of the recipient speaking it. Time delays represent a non-verbal instruction where the interventionist pauses an activity and waits for the recipient to respond. If the recipient does not respond after a few seconds, the interventionist models the expected response. Verbal instructions are expected to be presented at, or just above, the recipient's current communication level. For recipients that are non-verbal, this means instructions should be limited to one or two words.

For task variation, the interventionist is assessed based on how they vary instructions. This includes using different types of instructions to reinforce the same skill, or target speech, as well as interspersing mastered skills with target skills. Including skills the recipient has mastered, often called a maintenance skill, helps reinforce that skill to keep if from falling into disuse. It also helps keep the recipient motivated by a relatively easy activity in the midst of more difficult ones. This helps prevent frustration if the recipient is struggling with the target skills by providing an opportunity for success, access to the reinforcement, and praise from the interventionist.

Following the recipient's lead and identifying the natural reinforcer are related concepts. Part of naturalistic ABA methods is presenting learning opportunities in the context of an activity the recipient is interested in. For assessment, the interventionist should be observing the recipient in order to determine what activity they wish to engage in. After an activity is selected, the interventionist is expected to get involved in the activity to allow them to capture the recipient's attention and deliver an instruction. Capturing the recipient's attention involves identifying and controlling a natural reinforcer, often a toy or object involved in the activity, to hold or draw the recipient's focus.

Contingency is part of the consequence after the recipient has made a response. This has both a positive and a negative aspect depending on the response. In a positive scenario, the recipient has made an attempt at the target skill and the interventionist should deliver the reward immediately following the response to reinforce the behavior. In a negative scenario, the recipient has not made a responsible attempt and the interventionist is expected to withhold the reward, especially if the recipient begins engaging in disruptive behaviors.

Related to contingency as part of the consequence is the concept of rewarding attempts. To encourage the recipient and promote skill acquisition, the recipient should be rewarded for every reasonable attempt. A reasonable attempt is highly individualized and dependent on the recipient's current abilities. For instance, a recipient who is non-verbal may be rewarded for a communication skill attempt by gesturing or speaking a phoneme, where as a recipient with more verbal skills would need to speak the full word or phrase to be considered a reasonable attempt.

These categories are scored using a binary scale with the interventionist receiving a positive mark if they correctly demonstrated the technique. This limits the amount of feedback the interventionist receives on his or her performance in the video. Increasing the detail of the feedback would require significantly more time from the behavioral analyst scoring the probe. In research studies, it is common to have two analysts score each probe to mitigate misclassification. In practice, it is likely that only an analyst will review and provide feedback on the probes. Scoring the probes also means that there is a delay between when the interventionist records the video and when he or she receives feedback on implementation. This delay can prevent the interventionists from receiving the full benefit of the feedback. Studies have shown that immediately reviewing video of one's self implementing the interventions, along with feedback, helps the interventionists learn and feel more confident in their abilities [118].

An additional metric that is often recorded from the video probes when targeting communication skills is the verbal utterances of the recipient. This is often recorded in 10-to-15-second increments and may be categorized based on the instruction type the interventionist used to prompt the vocal attempt, or if it was a spontaneous vocalization. This metric usually focuses on in-context vocalization, not counting echolalic speech or disruptive behaviors.

8 New directions for incorporating technology

The research presented above illustrates some of the challenges faced by behavioral analysts providing adequate training and by non-clinical interventionists trying to learn and implement ABA treatments. Learning the treatment techniques requires access to education materials and training professionals. Although a large focus on in-person training can be seen in both the academic and the professional spheres, the logistical concerns of supporting individuals that are unable to attend intensive training courses have been scrutinized. To address location constraints, researchers have designed condensed courses [63] and implemented live teleconferencing [132]. These approaches do not address long-term support. Self-directed learning education modules provide training and reference materials but lack the interaction with trained professionals and access to feedback. Similar drawbacks exist with online educational platforms; however, there is the opportunity to create community features and keep information relevant that may help retain users for a long duration. What all of these approaches lack is long-term feedback. The duration of training programs often only

lasts weeks or months. During this time, it may be easy for the interventionist to gain fidelity in implementing the treatment on a specific set of goal; however, they may be unable determine new target skills or generalize the approach as the recipient improves. This could require the interventionist to have to seek out additional training sessions in order to continue to adapt the treatments.

In addition to addressing training challenges, technological designs need to emphasize the key benefits of the treatments. The research above illustrates the benefits that can be obtained when individuals utilize naturalistic ABA treatments with the child they interact with frequently. For the child, the studies show a greater improvement on social and communication assessments as well as improved affect. Likewise, the interventionists often report improved affect, reduced stress, and great confidence in their interactions with the child receiving the treatment. These benefits are what makes naturalistic methods effective. Technology brought in to enhance or support ABA training needs to be designed in regard to each benefit to ensure it is beneficial and utilized as a long-term solution.

Access to online educational materials for self-directed learning as discussed by [133] is an important step toward remote training of ABA methodologies and long-term support for practitioners. While this provides the information required to learn the approaches, it does not provide directed feedback that can be used to aid interventionists in personalizing the materials or build self-efficacy in implementation. Since assessment by a clinician is costly and may be impractical, technologies for multimedia processing can be utilized to reduce the cost of expert feedback though automated data collection processes.

The video probes currently used in naturalistic ABA treatment training provide the opportunity to use current multimedia processing research to gain insight into the interactions depicted along with reducing the time required by analysts to adequately score fidelity and provide feedback. Table 2 provides a brief overview of the areas of multimedia processing that could be utilized to extract information from video probes in regard to the current human-evaluation-based scoring methodologies. These scoring methodologies are multimodal and depend on both visual and auditory

Table 2 Naturalistic ABA evaluation criteria and relevant areas of technology that could be applied for automated analysis along with the expected feasibility based on current technologies

Evaluation category	Category	Relevant areas of technology	How it could be implemented	Feasibility (low/medium/high)
Opportunity to respond	Gaining attention	Attention classification	Identify dyadic poses that indicate attentive states	Medium
	Clear instruction	ASR, VAD, speaker separation, attention classification	Recognize and evaluate interventionist's instructions	High
Task variation	Instruction variation	ASR, VAD, speaker separation	Evaluate frequency and rate of alternation between forms of instructions	High
	Maintenance versus target skill	ASR, VAD, speaker separation	Analyze child's communication skills to determine target and maintenance tasks. Evaluate the parent's implementation to ensure proper balance	High
Contingency	Immediate reinforcement	VAD, speaker separation, object tracking, action detection	Identify recipient's vocal abilities and track reinforcement object passing to recipient	Medium/low
	Reinforcing earnest attempts	VAD, speaker separation, object tracking, action detection	Compare recipient's response to past responses to determine effort	Medium/low
Reinforcement	Following child's lead	Object tracking, activity detection	Analyze attention patterns and activity based on participant's poses	Medium/low
	Identifying natural reinforcer	Object tracking, activity detection	Identify important objects based on proximity to recipient and rate of interaction	Medium/low
Communication skills	Child responses	ASR, VAD, speaker separation	Identify and coordinate interventionist and recipient vocalizations to determine instructions, responses, and spontaneous speech	Medium

signals for proper assessment. Providing automated assessment requires examining techniques in video data and audio data processing. In video data processing research, object tracking, activity detection, and attention classification are relevant areas of study to this subject. Regarding audio data, voice activity detection, speaker separation, and automatic speech recognition (ASR) are applicable in order to extract verbal communication as well as vocalization attempts to evaluate the adult's instructions and the child's responses. These topics are discussed under the assumption that intervention sessions will follow the current standard, with no additional preparation required and utilizing ubiquitous recording devices.

9 Video image processing

9.1 Object tracking

Recognizing the activities depicted in the video probes requires identifying and tracking objects in the video. Tracking objects in images and videos involves discerning important areas of the frame from the background. For the PRT videos, there are two fundamental types of objects that need to be tracked: human participants, and toys and other objects involved in motivational activities. Tracking the participants and the objects needs to be handled differently. For the human participants, we need to be able to infer individual actions along with the interactions between each person. Object identification is relevant primarily in regard to its relationship to the child.

Tracking human figures in video frames can be accomplished using supervised learning methods [9]. Supervised learning techniques involve using known data to create models that can be used to infer knowledge about future data. In the case of PRT videos, it is assumed that the video has two human figures in each frame. The people in the video can then be tracked by using models that have been trained to detect human figures in images to identify where each individual is in the frame. This will allow the parent's and child's actions and interactions to be assessed throughout the video. Contrary to this, the objects in the video are dependent on the child and cannot be predicted. Identifying these objects requires using unsupervised learning techniques.

Unsupervised learning techniques rely on comparing unknown data in order to discover similarities and contrasts. For object detection in images, the task is to separate objects from the image background. This involves making inferences about saliency, often by looking at image contrast [39]. In video, changes between frames add an additional dimension to identifying objects. The color values of the pixels of neighboring frames are compared in order to determine areas of the video that are changing, indicating movement [66]. It is then presumed that moving objects of the video are important and garner the viewer's attention [121, 122].

PRT videos are a challenging medium for applying object tracking. As mentioned above, comparing the pixel transformation between frames is a key means for distinguishing important objects and identifying the same object in different frames. The algorithms often underperform in situations where there is a large amount of camera movement, or the objects being tracked in the video move too quickly or do not move at all. Both of these issues could be prevalent in PRT videos. Play activities may involve quick movements of the parent and child, or the individuals may rapidly move a toy. Additionally, as the videos are often recorded using a handheld device, the videos will exhibit some movement. Occlusion, or when the object is being obstructed from the camera by another object, is also a potential issue. This could be problematic in PRT videos as parents or children become obscured by objects, a book for example, or their bodies are not completely in frame. Likewise, important play objects, such as toys, may become obscured during play activities. Similar to occlusion, object deformation can be an issue for tracking algorithms. Although the algorithm may detect an object in one frame, it may fail to recognize the same object in a succeeding frame due to a different angle of the object being presented to the camera. Cluttered frames could also pose a problem for the segmentation tasks in object tracking. This is particularly challenging for models that use color contrast to differentiate foreground objects from the background of the image.

Tracking inanimate objects in the videos is mostly associated with PRT evaluation criteria regarding the reinforcer. Detecting the object the child is attending to could be a method for automatically determining what the natural reinforcer is in the situation. This can then be utilized along with information regarding the parent's activity in the same frame to determine whether the parent is following the child's lead or providing the child the reinforcer as part of the consequence step of PRT.

Tracking the participants in the videos is important for the majority of the evaluation categories for assessing parent implementation fidelity. In particular, inferring the activities of the human participants is essential to determining the child's state of attention, if the parent is following the child's lead and has identified the natural reinforcer, if the parent is providing appropriate reinforcement, and if the parent is providing a non-verbal instruction. To accomplish this, additional classification tasks need to be performed to extract information on the attention, activities being performed, and the dyadic interaction between the parent and child.

9.2 Activity detection and classification

Building on the concepts of object tracking, activity detection, and classification in video data relies on analyzing both spatial information about the configuration of people and objects in an individual frame, and temporal information relaying how those objects are transfigured in a sequence over time. This information is then used to infer the action or activity that is depicted in the video sequence. Numerous methodologies and technologies have been applied to the detection and classification of human activities in videos. These methods include identifying key frames in a sequence [99], organizing frames into a graph for analysis [12], examining sequences of identified images [76, 125], and exploring spatio-temporal representations of the individuals in a frame [26, 40]. Research on applying activity classification to individuals with autism has focused on single-person activities, primarily detecting self-stimulating or repetitive behaviors [16, 41, 47].

Activity classification for two or more people can take two common approaches, either the activities of each individual can be classified separately and then used to deduce a label about the entire scene [7, 19], or the individuals can be analyzed as single unit. One method for analyzing dyadic interactions is to create a single spatio-temporal graph of the major articulation points of each individual in the frame to use as date for training a classification model [128, 141].

Typically, research publications identify specific actions to classify from a given dataset, allowing for the use of supervised learning methods. In PRT implementation, the activities that the child and parent will engage in are likely not known prior to the session, and the specific activities are not important for evaluation. More generalized actions, such as whether or not the child is paying attention to the parent or if the parent has provided the reinforcer after the child makes a reasonable effort, are more important for evaluating fidelity. This simplifies the problem by allowing the system to be trained to recognize general poses or actions in the video, instead of classifying each activity that is depicted.

9.3 Attention and engagement classification

Detecting engagement and attention in the videos relies on poses or sequences that infer the individual state of focus. Methodologies for detecting attention rely on analyzing human head and body position and orientation. This has been used to estimate visual attention [6, 23, 135] and surmise social engagement [7, 105]. These studies used an exocentric perspective, with the camera not likely to be the object of attention. This means calculations of attention were independent of the camera location. Egocentric camera perspectives have also been used to infer attention. One example provides the caregiver with a wearable camera in order to capture periods of child engagement [96]. The child's attention to the caregiver can be inferred by periods when he or she is facing the camera. Attention and engagement are also important concepts in social robotics [25, 74, 102] and follow a similar methodology to classification techniques for human interactions.

Joint attention is an important concept for implementing PRT. Maintaining joint attention on a task will allow the parent to provide learning opportunities in a manner that is not disruptive to the child. To detect this in video data, the attention state of both individuals needs to be engaged with one another in a shared activity. From an exocentric perspective, this involves determining that the individuals in the video are attentive to one another or a shared object or task. As with activity and visual attention detection, current research into joint attention has focused on interpreting body, head, and facial orientation [95, 126]. In addition to this, rate of movement has also been determined to be diagnostic of attention [97, 98]. Intuitively, an individual attending to another individual would likely not be changing position or orientation rapidly. Preliminary work focusing on classifying joint attention in PRT video probes was undertaken in [37]. Their approach utilized spatio-temporal graphs of the parent and child that represented key facial and body locations, along with the individual's estimated visual attention.

Examining the research surrounding activity detection and classification is useful in automating evaluation of criteria based on the actions of the parent. These include following the child's lead, gaining the child's attention, and providing immediate reinforcement. The challenges of implementing activity classification in the PRT video probes are similar to object tracking problems. Successfully identifying the activity the individual is engaged in requires a clear depiction of the individual over successive frames. Partial or full occlusion, motion blur, or distortions in the depiction of the individual in the frame could lead to misclassifications of the activity or a classification not being possible. These issues could be addressed outside of the system by providing instructions on how to record the video probes. Within the system, predictive algorithms could be used to infer object locations within a frame based on its position in neighboring frames [37]. This is possible by leveraging the domain knowledge that the frame should depict two individuals.

10 Audio processing

10.1 Voice activity detection and speaker separation

Voice activity detection (VAD) encompasses the preprocessing techniques for discriminating speech signals from other

noises in an audio file. Generally, approaches to classifying speech and nonspeech signals involve using discriminatory feature sets, statistical approaches, or machine learning techniques [140]. A common feature-based technique is the use of frequency ranges as a filter for selecting speech signals [1, 79]. Statistical approaches focus on modeling the noise spectra using a defined distribution in order to extract impertinent signals [11]. In addition to these methods, both unsupervised and supervised learning techniques have been explored [21, 34, 100, 109].

Much of the attention of the research community is devoted to extracting noise signals from the environment that may obscure speech. For PRT evaluation, it is more important to identify areas of speech versus noise. If the instruction is obscured by noise, it would not achieve the criteria for a clear instruction. It is problematic for recording responses and spontaneous speech from the child that may occur during periods obscured by environmental noise [48, 73].

After noise is filtered out, speech signals need to be attributed to either the child or the parent. The methodology for separating speakers is similar to VAD and can be accomplished using feature, statistical, or machine learning-based approaches. Research conducted on PRT video probes determined that noise, parent speech, and child vocalizations could be adequately differentiated using supervised machine learning [38].

In evaluating PRT videos, this separation becomes an important component of determining the type of speech events that has occurred. Ideally, PRT audio should capture a parent's instructions, followed by a child's response, and may be finalized with a verbal assurance in conjunction with the parent providing the reinforcer as part of the immediate consequences criteria. In addition to this pattern, spontaneous speech from the child would be important to note for feedback and long-term communication skill progression tracking. These speech events could also be correlated with attention and activity classifications to determine whether the adult instruction was provided at an ideal time to capitalize on the child's engagement.

10.2 Automatic speech recognition

Automatic speech recognition (ASR) systems have become a ubiquitous feature of many modern applications. For the most part, these systems function adequately for the majority of users. Analyzing speech involves receiving the sound wave as a time-series signal, either transforming the signal to isolate discriminative features for classification [17] or processing the raw signal through a trained classification model. Identifying words in the signal can be accomplished by isolating and classifying phonemes, then using a lexicon to construct words [33], or training the model to recognize the signal pattern of full words [10, 101, 142].

Current commercially available ASRs have adopted deep learning as core components or their systems. Microsoft Cognitive Toolkit [81], Google Speech to Text [87], and Houndify [113] are a sample of the products that are available utilizing neural networks to improve automated speech tasks. In addition to providing improved speech detection, these systems aim to provide more robust solutions. A predominant issue the current generation of ASRs is focusing on, and one that is important to speech recognition in PRT video probes, is filtering audio to remove environmental noise not pertaining to the speech signals.

Performing speech recognition on children presents additional challenges. At an auditory level, children's voices tend to be higher frequency and display more rational and spectral variability [72]. Regarding language modeling, children are more prone to mispronouncing words more than adults, have a restricted vocabulary, and tend to speak at a lower rate [93]. These challenges are more apparent the younger the child is. Adult speech has been used to improve the recognition of child speech by using contextual information from adult speech in a dyadic conversation [67], or by utilizing adult speech samples when training classification models [75, 111]. In addition to detection research, studies involving children have examined native language detection and analysis [8], gender and a naturalized speech [49], and detecting speech pathologies [22, 111].

Much of the research regarding the implementation of ASR systems for individuals with autism has focused on diagnosis and emotion detection. Looking at autism detection, [137, 138] used the Language Environment Analysis (LENA) audio recording system to record children with autism in a home environment. Their goal was to alleviate the human process time for evaluating language skills for people with autism. The LENA recording system was also used by [86] to analyze vocalizations of children with autism, and their interaction with adults. Deep learning techniques have also been explored for diagnosing autism [18].

Exploration of the application of ASRs for emotion detection in children with autism was undertaken by [78]. For this work, the researchers examined and compared participants from different nationalities and included children with or without an autism diagnosis. They found that the system had a higher detection recall rate for the children without an ASD diagnosis.

In addition to recognizing the speech, extracting the semantic meaning of the adult's instructions is important for evaluating implementation fidelity. Semantic parsing of instructions is part of natural language processing (NLP) research and is common in human–computer and human–robot interaction. In these studies, the objective is to parse the human instruction into a set of known words,

primitives, related to various actions [5, 42, 134]. Different methods have been used for semantic parsing including unsupervised clustering techniques [91], SVM [94], deep learning [30, 139], and logic approaches [106].

For evaluating PRT video probes, an ASR system needs to be able to recognize the parent's instructions along with the child's vocalizations. This is important for determining whether instructions are at the child's level of speech, and examining task variation between both instruction types, and the ratio of target to maintenance skills. The majority of currently available speech recognition systems should be capable of analyzing parent speech in the videos. However, this is made more difficult by the parents often exhibiting speech patterns that are not typical of adult conversational speech. The PRT videos depict parents using child-directed speech, or baby talk, to engage the child. This is exemplified by using a higher tone of voice, exaggerating emotion and excitement, and elongating syllables, particularly at the end of words. This could be problematic during speaker separation tasks as the higher pitch of the parent's voice could incorrectly be attributed as a child's voice. Differences in pronunciations could affect an ASR's ability to ascertain the appropriate words. These systems are trained with adult speech that may not encompass the word transformations, resulting in incorrectly recognized text.

In addition to the difficulties associated with child speech recognition, examining the speech and vocalizations of children with autism could provide more challenges depending on the child's vocal skills. In PRT, an adequate attempt at speech is determined by the child's vocal ability. For children that exhibit non-verbal behavior or only speech in one or two word phrases, an adequate attempt at target skills does not need to consist of pronouncing a full word. In these instances, vocalizing the starting phoneme of the desired word, or any vocalization at all, could be significant enough to elicit the reinforcer. Because of this, it is important to be able to identify child vocalizations at the phoneme level. In addition to phoneme-based ASR systems, relevant research into nonspeech vocal events [31, 32] could be relevant in classification of single phoneme vocalizations.

11 Current feasibility of automating evaluation assessments

The research presented above illustrates successful approaches for addressing individual tasks such as object tracking, activity detection, or speech recognition. The evaluation criteria often require multimodal analysis for adequate evaluation. Given the current state of technology, the feasibility of successfully automated detection for the criteria varies depending on the modalities involved and level of subjectivity. The most likely criteria to be successfully automated are clear instruction, instruction variation, and maintenance vs target skill. These categories are based on analysis of the interventionist's speech. Although child-directed speech patterns make recognition more difficult, instructions in PRT are expected to be direct and reflect the language level of the recipient. Current ASR systems could likely extract the adult speech, and ASR systems could be refined using labeled child-directed speech to become more robust. The instructions should not be complicated sentences, which makes modern NLP techniques adequate for parsing instructions. Reducing the instruction to a particular phrase form would allow the system to determine whether there is sufficient variation in the instructions. Evaluating whether the instructions are at the recipient's speaking level would require supplying *a priori* information to the system, or allowing it to assess the recipient's ability over time. Evaluation criteria involving the recipient's vocalizations will be more challenging to assess than the interventionist's. This is largely due to concerns involving the detection of nonspeech vocalizations, intelligibility, and the general challenges with detecting child speech.

Evaluating immediate reinforcement and reinforcing earnest attempts would require object tracking, human activity classification, and speech analysis to be successfully assessed. Under certain scenarios, this could be relatively straightforward. If the interventionist has control over an object the recipient is motivated by, evaluating reinforcement could be based on tracking the object passing from the interventionist to the recipient. This interaction could be assessed based on whether or not it occurred in a timely manner after a response, and if that response was considered adequate based on information regarding the recipient's ability. This will be more complicated to assess whether the reinforcement is the continuation of an activity, or whether the dyad are engaged in a shared activity. These instances will rely on human activity detection. Basing the assessment on detecting phrases praising the recipient's performance may be an alternative approach that could make classification more robust.

Following the child's lead and identifying the natural reinforcer are also based primarily on object tracking and dyadic activity recognition. Correct assessment of these categories involves the interventionist recognizing the object or activity the recipient is motivated by and then integrating himself or herself into it. Evaluation would be based on how the individuals interact between each other and the motivational object. Inference would likely rely on proximity between the individuals. This could be problematic in two-dimensional space when addressing the camera perspective. If the interventionist is standing behind an object the child is interacting with, but not involving himself or herself in the interactions, this could be classified as a false positive.

Unlike the other categories that assess activity, this criterion examines human behavior. This could be problematic as it is dependent on visual cues of attention. Different individuals, particularly a child with autism, may not exhibit outward signs of attention, making classification more difficult. Additionally, classification of attention is more subjective than identifying specific activities. Unlike activities that rely on a structured series of events, attention can be surmised based on a limited number of visual cues. This could allow attention classification to be more generalized. As with the previous categories, in simple scenarios where the interventionist gains control of an object, and the recipient is motivated to engage with, the interaction may not be difficult to classify. In this instance, attention can be inferred by determining whether the child is looking at the interventionist, or the object in his or her control, and the recipient is not engaged in a separate activity. Periods of shared attention will be more difficult to classify depending on the activity.

12 Beyond current practices

The discussion of the application of technology to analyze video probes is based on the assumption that no changes will be made to the procedures currently being used in recording the videos, the environment where interactions are taking place, and the devices being utilized. This is based on the idea that using ubiquitous devices and limiting required preparations afford the interventionist the ability to initiate sessions naturally and spontaneously. However, it is valuable to explore how incorporating new devices would aid in automated assessment. In particular, automated evaluations would benefit from the utilization of new camera technology, audio recording equipment, and marked or enhanced objects.

Three-dimensional cameras utilize two lenses to capture a stereoscopic image that provides information regarding depth in addition to pixel color values. Adding the depth data to the image aids detection and classification of human actions [41, 107, 144], improved occlusion handling [88], and aids in estimating visual focus [135]. Having perspective in the image would aid in detecting the visual attention of the child. In particular, this would provide improved estimates of visual focus when the interventionist and an object are overlapping.

Speaker separation was described as a challenge facing automated audio analysis in the video probes. Under the current assumption, this would be collected using a single handheld device. Incorporating wearable microphones would afford the opportunity to improve data collection and have different streams for each speaker. The LENA system [86, 110, 137, 138] employed discrete wearable microphones for the child. Using this approach on both individuals in the interaction would enable an automated processing to separate individual speakers based on the strength of the signal from their assigned devices. This would also provide perspective on how environmental sounds affect each individual, which could be used to infer when ambient noise could be a distraction.

Under the philosophy of naturalistic ABA, any object a recipient selects could be used as a reinforcer. In practice, it is likely that the interventionist could promote the recipient selecting specific objects. This would generally be enacted when the recipient knows a favored object of the recipient and places the object in the environment before a session. This is often utilized by placing the object in a location that is visible, but unreachable to the recipient. The recipient will then need to initiate communication to acquire the object. Under this practice, the object could be marked or enhanced to aid in tracking the object in the video. Using retroreflective markers was a common approach to early object tracking [20]. Utilizing common color patterns, similar to image codes [80], could allow for improved object identification if the tracking system was pre-trained on the patterns.

Embedding sensors into the object to allow it to transmit data would aid in object tracking and identification. By attaching inertial sensors and using wireless communication, the automated system could detect when an object is moved, providing additional information for tracking [143]. During the session, this would likely indicate that the object is being manipulated by one of the dyads. When added to visual information, this could be used to make inferences regarding detecting attention, identifying the natural reinforcer, and providing immediate consequences. Radio-frequency identification (RFID) tags should be explored to determine whether they could be used for localized tracking. This could provide a cost-effective solution for attaching sensors to objects.

13 Conclusions and areas for future work

Although ABA techniques, particularly naturalistic methodologies, have been shown to be effective treatments for aiding children with autism to develop social and communication skills, the scarcity of resources and time required to train interventionists provides a significant barrier. This barrier can be particularly difficult to overcome for parents that may live in rural locations or are unable to cover the costs of enrolling in training programs. Even after receiving training, parents often have little support to help them adapt the treatments to target new skills. The application of current multimedia processing tools could be an effective means for providing more support and feedback to parents of children with autism that are interested in using behavioral treatments to improve the child's vocal abilities. Utilizing these technologies could alleviate the

amount of time needed for clinicians to provide feedback, which would result in more fulfilling learning experiences. Additionally, embracing technological solutions would aid in disseminating information and facilitate distance learning. This would ensure that parents have greater access to learning materials and consultations with trained professionals.

The research presented in this paper illustrated how current areas of technology could be directly applied to naturalistic ABA evaluation metrics; however, this represents a foundation with several immediate and future goals left to pursue. The immediate goals focus primarily on evaluating the classification techniques outlined in this paper. Preliminary work has been presented on detecting attention [37] and speaker separation [38]. After applying application models, how the data are presented to the clinician and ultimately to the parent interventionist needs to be explored in greater detail. With an associated human cost, automatically assessing data provides the opportunity to increase the amount of information that is collected by looking at finer resolutions of time. Current assessments evaluate users at one- or two-minute intervals; however, this could be reduced in an automated system. The types of metrics that are provided and the data visualizations utilized in their presentation need to be developed in conjunction with behavior analysts. This will aid in determining important metrics and prevent overwhelming users with superfluous data.

Along with increasing the metrics, technology could provide a means for measuring aspects of PRT implementation that are not currently trackable. Research into emotion detection [84, 92] could provide a means of tracking the effect PRT training is having on the parent and child's affective state. This, along with metrics on progress and expert evaluation, could be utilized to create a training tool that would motivate parents to continue to practice and improve in teaching verbal communication skills to their children. Pursuing the detection of affect and engagement are important future goals.

The ultimate goal is creating a feedback system that supports parents in the early learning stages and over the long term by aiding in adapting implementation practices as their child's skills continue to grow. Long-term usage of a system designed for supporting parent implementation of PRT training needs to incorporate features that will motivate continual use of the application. Self-regulatory learning provides a framework that can be utilized in the design of the software to foster self-motivation. Design features aid the user in creating goals and monitoring achievements to facilitate value in the treatment and use of the software. Maintaining a connection with a clinician through the application would allow the expert to help the user set goals, overcome plateaus, and provide social pressure for continued use [50].

Acknowledgements The authors thank Arizona State University and the National Science Foundation for their funding support. This material is partially based upon work supported by the National Science Foundation under Grant Nos. 1069125 and 1828010.

References

1. Aneeja, G., Yegnanarayana, B.: Single frequency filtering approach for discriminating speech and nonspeech. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **23**(4), 705–717 (2015)
2. Baer, D.M., Wolf, M.M., Risley, T.R.: Some current dimensions of applied behavior analysis. J. Appl. Behav. Anal. **1**(1), 91–97 (1968)
3. Bagaiolo, L.F., Mari, JdJ, Bordini, D., Ribeiro, T.C., Martone, M.C.C., Caetano, S.C., Brunoni, D., Brentani, H., Paula, C.S.: Procedures and compliance of a video modeling applied behavior analysis intervention for brazilian parents of children with autism spectrum disorders. Autism **21**, 603–610 (2017)
4. Baker-Ericzén, M.J., Stahmer, A.C., Burns, A.: Child demographics associated with outcomes in a community-based pivotal response training program. J. Positive Behav. Interv. **9**(1), 52–60 (2007)
5. Bastianelli, E., Castellucci, G., Croce, D., Basili, R., Nardi, D.: Effective and robust natural language understanding for human-robot interaction. In: Proceedings of the Twenty-first European Conference on Artificial Intelligence, pp. 57–62. IOS Press (2014)
6. Baxter, R.H., Leach, M.J., Mukherjee, S.S., Robertson, N.M.: An adaptive motion model for person tracking with instantaneous head-pose features. IEEE Signal Process. Lett. **22**(5), 578–582 (2015)
7. Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., Murino, V.: Social interactions by visual focus of attention in a three-dimensional environment. Exp. Syst. **30**(2), 115–127 (2013)
8. Boril, H., Zhang, Q., Ziaei, A., Hansen, J.H., Xu, D., Gilkerson, J., Richards, J.A., Zhang, Y., Xu, X., Mao, H., others: Automatic assessment of language background in toddlers through phonotactic and pitch pattern modeling of short vocalizations. In: WOCCI, pp. 39–43 (2014)
9. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
10. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE (2016)
11. Chang, J.H., Kim, N.S., Mitra, S.K.: Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. **54**(6), 1965–1976 (2006)
12. Chen, C.Y., Grauman, K.: Efficient activity detection in untrimmed video with max-subgraph search. IEEE Trans. Pattern Anal. Mach. Intell. **39**(5), 908–921 (2017)
13. Children's Hospital at Sacred Heart: Children's hospital at sacred heart—autism center https://sacred-heart.org/childrenshospital/main/services/page/?id=1002. Acessed 19 May 2018 (2018)
14. Choice Autism Center: Our programs. https://choiceautismcenter.com/our-programs/. Accessed 19 May 2018 (2018)
15. Coolican, J., Smith, I.M., Bryson, S.E.: Brief parent training in pivotal response treatment for preschoolers with autism. J. Child Psychol. Psychiatry **51**(12), 1321–1330 (2010)
16. Coronato, A., De Pietro, G., Paragliola, G.: A situation-aware system for the detection of motion disorders of patients with

autism spectrum disorders. Exp. Syst. Appl. **41**(17), 7868–7877 (2014)
17. Dave, N.: Feature extraction methods LPC, PLP and MFCC in speech recognition. Int. J. Adv. Res. Eng. Technol. **1**(6), 1–4 (2013)
18. Deng, J., Cummins, N., Schmitt, M., Qian, K., Ringeval, F., Schuller, B.: Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. In: Proceedings of the 2017 International Conference on Digital Health, pp. 53–57. ACM (2017)
19. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4772–4781 (2016)
20. Dorfmüller, K.: Robust tracking for augmented reality using retroreflective markers. Comput. Gr. **23**(6), 795–800 (1999)
21. Drugman, T., Stylianou, Y., Kida, Y., Akamine, M.: Voice activity detection: merging source and filter-based information. IEEE Signal Process. Lett. **23**(2), 252–256 (2016)
22. Dudy, S., Bedrick, S., Asgari, M., Kain, A.: Automatic analysis of pronunciations for children with speech sound disorders. Comput. Speech Lang. **50**, 62–84 (2017)
23. Duffner, S., Garcia, C.: Visual focus of attention estimation with unsupervised incremental learning. IEEE Trans. Circuits Syst. Video Technol. **26**(12), 2264–2272 (2016)
24. Estes, A., Vismara, L., Mercado, C., Fitzpatrick, A., Elder, L., Greenson, J., Lord, C., Munson, J., Winter, J., Young, G.: The impact of parent-delivered intervention on parents of very young children with autism. J. Autism Dev. Disord. **44**(2), 353–365 (2014)
25. Foster, M.E., Gaschler, A., Giuliani, M.: How can i help you': comparing engagement classification strategies for a robot bartender. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 255–262. ACM (2013)
26. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4346–4354 (2015)
27. Gengoux, G.W., Berquist, K.L., Salzman, E., Schapp, S., Phillips, J.M., Frazier, T.W., Minjarez, M.B., Hardan, A.Y.: Pivotal response treatment parent training for autism: findings from a 3-month follow-up evaluation. J. Autism Dev. Disord. **45**(9), 2889–2898 (2015)
28. Gillesen, J.C., Barakova, E., Huskens, B.E., Feijs, L.M.: From training to robot behavior: Towards custom scenarios for robotics in training programs for ASD. In: 2011 IEEE International Conference on Rehabilitation Robotics (ICORR), pp. 1–7. IEEE (2011)
29. Gillett, J.N., LeBlanc, L.A.: Parent-implemented natural language paradigm to increase language and play in children with autism. Res. Autism Spec. Disord. **1**(3), 247–255 (2007)
30. Goldberg, Y.: Neural network methods for natural language processing. Synth. Lect. Human Lang. Technol. **10**(1), 1–309 (2017)
31. Gosztolya, G.: Detecting laughter and filler events by time series smoothing with genetic algorithms. In: International Conference on Speech and Computer, pp. 232–239 (2016)
32. Gosztolya, G., Grósz, T., Busa-Fekete, R., Tóth, L.: Determining native language and deception using phonetic features and classifier combination. Interspeech (2016)
33. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (icassp), pp. 6645–6649. IEEE (2013)
34. Górriz, J.M., Ramírez, J., Lang, E.W., Puntonet, C.G.: Hard c-means clustering for voice activity detection. Speech Commun. **48**(12), 1638–1649 (2006)
35. Hardan, A.Y., Gengoux, G.W., Berquist, K.L., Libove, R.A., Ardel, C.M., Phillips, J., Frazier, T.W., Minjarez, M.B.: A randomized controlled trial of pivotal response treatment group for parents of children with autism. J. Child Psychol. Psychiatry **56**(8), 884–892 (2015)
36. Harper, C.B., Symon, J.B., Frea, W.D.: Recess is time-in: using peers to improve social skills of children with autism. J. Autism Dev. Disord. **38**(5), 815–826 (2008)
37. Heath, C.D., Venkateswara, H., McDaniel, T., Panchanathan, S.: Detecting attention in pivotal response treatment video probes. In: International Conference on Smart Multimedia (2018)
38. Heath, C.D., McDaniel, T., Venkateswara, H., Panchanathan, S.: Parent and child voice activity detection in pivotal response treatment video probes. In: Human Computer Interaction International (2019)
39. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
40. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-RNN: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5308–5317 (2016)
41. Jazouli, M., Elhoufi, S., Majda, A., Zarghili, A., Aalouane, R.: Stereotypical motor movement recognition using microsoft kinect with artificial neural network. World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. **10**(7), 1270–1274 (2016)
42. Jernite, Y., Srinet, K., Gray, J., Szlam, A.: CraftAssist instruction parsing: semantic parsing for a minecraft assistant. Preprint arXiv:1905.01978 (2019)
43. Johnson, N., Frenn, M., Feetham, S., Simpson, P.: Autism spectrum disorder: parenting stress, family functioning and health-related quality of life. Fam. Syst. Health **29**(3), 232 (2011)
44. Jones, E.A., Feeley, K.M.: Parent implemented joint attention intervention for preschoolers with autism. J. Speech Lang. Pathol. Appl. Behav. Anal. **4**(1), 74–89 (2009). https://doi.org/10.1037/h0100251
45. Kane, M., Connell, J.E., Pellecchia, M.: A quantitative analysis of language interventions for children with autism. Behav. Anal. Today **11**(2), 128 (2010)
46. Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., Berry, K.: Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. J. Consult. Clin. Psychol. **83**(3), 554 (2015)
47. Khan, N.A., Sawand, M.A., Qadeer, M., Owais, A., Junaid, S., Shahnawaz, P.: Autism detection using computer vision. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **17**(4), 256 (2017)
48. Kim, J., Hahn, M.: Voice activity detection using an adaptive context attention model. IEEE Signal Process. Lett. **25**(8), 1181 (2018)
49. Kim, J., Englebienne, G., Truong, K., Evers, V.: Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning. Interspeech (2017)
50. Kitsantas, A., Kavussanu, M.: Acquisition of sport knowledge and skill. In: Zimmerman, B., Schunk, D. (eds.) Handbook of Self-regulation of Learning and Performance, pp. 217–233. Routledge, New York, London (2011)
51. Koegel, L.K., Camarata, S.M., Valdez-Menchaca, M., Koegel, R.L.: Setting generalization of question-asking by children with autism. Am. J. Mental Retard. **102**(4), 346–357 (1997)
52. Koegel, L.K., Koegel, R.L., Harrower, J.K., Carter, C.M.: Pivotal response intervention i: overview of approach. J. Assoc. Persons Severe Handicaps **24**(3), 174–185 (1999)

53. Koegel, L.K., Koegel, R.L., Shoshan, Y., McNerney, E.: Pivotal response intervention II: preliminary long-term outcome data. J. Assoc. Persons Severe Handicaps **24**(3), 186–198 (1999)
54. Koegel, L.K., Carter, C.M., Koegel, R.L.: Teaching children with autism self-initiations as a pivotal response. Top. Lang. Disord. **23**(2), 134–145 (2003)
55. Koegel, L.K., Koegel, R.L., Green-Hopkins, I., Barnes, C.C.: Brief report: question-asking and collateral language acquisition in children with autism. J. Autism Dev. Disord. **40**(4), 509–515 (2010)
56. Koegel, L.K., Singh, A.K., Koegel, R.L., Hollingsworth, J.R., Bradshaw, J.: Assessing and improving early social engagement in infants. J. Positive Behav. Interv. **16**(2), 69–80 (2014)
57. Koegel, R.L., Schreibman, L., Good, A., Cerniglia, L., Murphy, C., Koegel, L.: How to teach pivotal behaviors to children with autism: a training manual. University of California, Santa Barbara (1988)
58. Koegel, R.L.: A natural language teaching paradigm for nonverbal autistic children. J. Autism Dev. Disord. **17**(2), 187–200 (1987)
59. Koegel, R.L., O'Dell, M., Dunlap, G.: Producing speech use in nonverbal autistic children by reinforcing attempts. J. Autism Dev. Disord. **18**(4), 525–538 (1988)
60. Koegel, R.L., Koegel, L.K., Surratt, A.: Language intervention and disruptive behavior in preschool children with autism. J. Autism Dev. Disord. **22**(2), 141–153 (1992)
61. Koegel, R.L., Bimbela, A., Schreibman, L.: Collateral effects of parent training on family interactions. J. Autism Dev. Disord. **26**(3), 347–359 (1996)
62. Koegel, R.L., Camarata, S., Koegel, L.K., Ben-Tall, A., Smith, A.E.: Increasing speech intelligibility in children with autism. J. Autism Dev. Disord. **28**(3), 241–251 (1998)
63. Koegel, R.L., Symon, J.B., Kern Koegel, L.: Parent education for families of children with autism living in geographically distant areas. J. Positive Behav. Interv. **4**(2), 88–103 (2002)
64. Koegel, R.L., Vernon, T.W., Koegel, L.K.: Improving social initiations in young children with autism using reinforcers with embedded social interactions. J. Autism Dev. Disord. **39**(9), 1240–1251 (2009)
65. Koegel, R.L., Bradshaw, J.L., Ashbaugh, K., Koegel, L.K.: Improving question-asking initiations in young children with autism using pivotal response treatment. J. Autism Dev. Disord. **44**(4), 816–827 (2014)
66. Koh, Y.J., Kim, C.S.: Primary object segmentation in videos based on region augmentation and reduction. In: CVPR, vol. 1, p. 7 (2017)
67. Kumar, M., Bone, D., McWilliams, K., Williams, S., Lyon, T.D., Narayanan, S.: Multi-scale context adaptation for improving child automatic speech recognition in child-adult spoken interactions. Proc. Interspeech **2017**, 2730–2734 (2017)
68. Laski, K.E., Charlop, M.H., Schreibman, L.: Training parents to use the natural language paradigm to increase their autistic children's speech. J. Appl. Behav. Anal. **21**(4), 391–400 (1988)
69. Lawton, K., Kasari, C.: Teacher-implemented joint attention intervention: pilot randomized controlled study for preschoolers with autism. J. Consul. Clin. Psychol. **80**(4), 687 (2012)
70. Leaf, J.B., Leaf, R., McEachin, J., Taubman, M., Ala'i-Rosales, S., Ross, R.K., Smith, T., Weiss, M.J.: Applied behavior analysis is a science and therefore, progressive. J. Autism Dev. Disord. **46**(2), 720–731 (2016)
71. Lecavalier, L., Smith, T., Johnson, C., Bearss, K., Swiezy, N., Aman, M.G., Sukhodolsky, D.G., Deng, Y., Dziura, J., Scahill, L.: Moderators of parent training for disruptive behaviors in young children with autism spectrum disorder. J. Abnormal Child Psychol. **45**(6), 1235–1245 (2017)
72. Lee, S., Potamianos, A., Narayanan, S.: Acoustics of children's speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. **105**(3), 1455–1468 (1999)
73. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 745–777 (2014)
74. Li, L., Xu, Q., Tan, Y.K.: Attention-based addressee selection for service and social robots to interact with multiple persons. In: Proceedings of the Workshop at SIGGRAPH Asia, pp. 131–136. ACM (2012)
75. Liao, H., Pundak, G., Siohan, O., Carroll, M.K., Coccaro, N., Jiang, Q.M., Sainath, T.N., Senior, A., Beaufays, F., Bacchiani, M.: Large vocabulary automatic speech recognition for children. In: Sixteenth Annual Conference of the International Speech Communication Association, pp. 1611–1615 (2015)
76. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in lstms for activity detection and early detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1942–1950 (2016)
77. Machalicek, W., O'Reilly, M.F., Rispoli, M., Davis, T., Lang, R., Franco, J.H., Chan, J.M.: Training teachers to assess the challenging behaviors of students with autism using video teleconferencing. Educ. Train. Autism Dev. Disabil. **45**, 203–215 (2010)
78. Marchi, E., Schuller, B., Baron-Cohen, S., Golan, O., Bölte, S., Arora, P., Häb-Umbach, R.: Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages. In: Sixteenth Annual Conference of the International Speech Communication Association, pp. 115–119 (2015)
79. McLoughlin, I.V.: The use of low-frequency ultrasound for voice activity detection. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
80. Mehner, W., Boltes, M., Mathias, M., Leibe, B.: Robust marker-based tracking for measuring crowd dynamics. In: International Conference on Computer Vision Systems, pp. 445–455. Springer (2015)
81. Microsoft: Microsoft cognitive toolkit (CNTK), an open source deep-learning toolkit https://github.com/microsoft/cntk. Accessed 24 June 2018 (2018)
82. Mohammadzaheri, F., Koegel, L.K., Rezaee, M., Rafiee, S.M.: A randomized clinical trial comparison between pivotal response treatment (PRT) and structured applied behavior analysis (ABA) intervention for children with autism. J. Autism Dev. Disord. **44**(11), 2769–2777 (2014)
83. Mohammadzaheri, F., Koegel, L.K., Rezaei, M., Bakhshi, E.: A randomized clinical trial comparison between pivotal response treatment (PRT) and adult-driven applied behavior analysis (ABA) intervention on disruptive behaviors in public school children with autism. J. Autism Dev. Disord. **45**(9), 2899–2907 (2015)
84. Naim, I., Tanveer, M.I., Gildea, D., Hoque, M.E.: Automated prediction and analysis of job interview performance: the role of what you say and how you say it. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–6. IEEE (2015)
85. Nefdt, N., Koegel, R., Singer, G., Gerber, M.: The use of a self-directed learning program to provide introductory training in pivotal response treatment to parents of children with autism. J. Positive Behav. Interv. **12**(1), 23–32 (2010)
86. Pawar, R., Albin, A., Gupta, U., Rao, H., Carberry, C., Hamo, A., Jones, R.M., Lord, C., Clements, M.A.: Automatic analysis of LENA recordings for language assessment in children aged five to fourteen years with application to individuals with autism. In: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 245–248. IEEE (2017)

87. Perrier, A.: Google upgrades its speech-to-text service with tailored deep-learning models https://www.infoq.com/news/2018/05/google-speech-to-text-api. Accessed 24 June 2018 (2018)
88. Pi, J., Gu, Y., Hu, K., Cheng, X., Zhan, Y., Wang, Y.: Real-time scale-adaptive correlation filters tracker with depth information to handle occlusion. J. Electron. Imag. **25**(4), 043022 (2016)
89. Pierce, K., Schreibman, L.: Increasing complex social behaviors in children with autism: effects of peer-implemented pivotal response training. J. Appl. Behav. Anal. **28**(3), 285–295 (1995)
90. Pierce, K., Schreibman, L.: Multiple peer use of pivotal response training to increase social behaviors of classmates with autism: results from trained and untrained peers. J. Appl. Behav. Anal. **30**(1), 157–160 (1997)
91. Poon, H., Domingos, P.: Unsupervised semantic parsing. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: vol. 1, pp. 1–10. Association for Computational Linguistics (2009)
92. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174**, 50–59 (2016)
93. Potamianos, A., Narayanan, S.: Spoken dialog systems for children. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, vol. 1, pp. 197–200. IEEE (1998)
94. Pradhan, S.S., Ward, W.H., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow semantic parsing using support vector machines. Proc. Human Lang. Technol. Conf. North Am. Chapter Assoc. Comput. Linguist. HLT-NAACL **2004**, 233–240 (2004)
95. Presti, L., Sclaroff, S., Rozga, A.: Joint alignment and modeling of correlated behavior streams. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 730–737 (2013)
96. Pusiol, G., Soriano, L., Frank, M.C., Fei-Fei, L.: Discovering the signatures of joint attention in child-caregiver interaction. In: Proceedings of the Cognitive Science Society, vol. 36 (2014)
97. Rajagopalan, S.S., Murthy, O.R., Goecke, R., Rozga, A.: Play with me-measuring a child's engagement in a social interaction. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
98. Rajagopalan, S.S., Morency, L.P., Baltrusaitis, T., Goecke, R.: Extending long short-term memory for multi-view structured learning. In: European Conference on Computer Vision, pp. 338–353. Springer (2016)
99. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2650–2657 (2013)
100. Sadjadi, S.O., Hansen, J.H.: Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. **20**(3), 197–200 (2013)
101. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4580–4584. IEEE (2015)
102. Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W., Paiva, A.: Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: 2011 6th ACM/IEEE International Conference on Human–Robot Interaction (HRI), pp. 305–311. IEEE (2011)
103. Schreibman, L., Kaneko, W.M., Koegel, R.L.: Positive affect of parents of autistic children: a comparison across two teaching techniques. Behav. Ther. **22**(4), 479–490 (1991)
104. Schreibman, L., Dawson, G., Stahmer, A.C., Landa, R., Rogers, S.J., McGee, G.G., Kasari, C., Ingersoll, B., Kaiser, A.P., Bruinsma, Y.: others: Naturalistic developmental behavioral interventions: empirically validated treatments for autism spectrum disorder. J. Autism Dev. Disord. **45**(8), 2411–2428 (2015)
105. Sener, F., Ikizler-Cinbis, N.: Two-person interaction recognition via spatial multiple instance embedding. J. Vis. Commun. Image Represent. **32**, 63–73 (2015)
106. Shaalan, K.: Extending prolog for better natural language analysis. In: Proceeding of the 1st Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), pp. 225–236 (2019)
107. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
108. Sherer, M.R., Schreibman, L.: Individual behavioral profiles and predictors of treatment effectiveness for children with autism. J. Consul. Clin. Psychol. **73**(3), 525 (2005)
109. Shin, J.W., Chang, J.H., Kim, N.S.: Voice activity detection based on statistical models and machine learning approaches. Comput. Speech Lang. **24**(3), 515–530 (2010)
110. Shivakumar, S.S., Loeb, H., Bogen, D.K., Shofer, F., Bryant, P., Prosser, L., Johnson, M.J.: Stereo 3d tracking of infants in natural play conditions. In: 2017 International Conference on Rehabilitation Robotics (ICORR), pp. 841–846. IEEE (2017)
111. Smith, D., Sneddon, A., Ward, L., Duenser, A., Freyne, J., Silvera-Tawil, D., Morgan, A.: Improving child speech disorder assessment by incorporating out-of-domain adult speech. Proc. Interspeech **2017**, 2690–2694 (2017)
112. Smith, I.M., Flanagan, H.E., Garon, N., Bryson, S.E.: Effectiveness of community-based early intervention based on pivotal response treatment. J. Autism Dev. Disord. **45**(6), 1858–1872 (2015)
113. SoundHound Inc.: Houndify by SoundHound inc. https://houndify.com/. Accessed 2 Feb 2019 (2019)
114. Southwest Autism Research and Resouce Center: Parents and caregivers https://autismcenter.org/parents-and-caregivers. Accessed 19 May 2018 (2016)
115. Stahmer, A.C.: Teaching symbolic play skills to children with autism using pivotal response training. J. Autism Dev. Disord. **25**(2), 123–141 (1995)
116. Stahmer, A.C., Schreibman, L., Powell, N.P.: Social validation of symbolic play training for children with autism. J. Early Intensive Behav. Interv. **3**(2), 196 (2006)
117. Steiner, A.M., Gengoux, G.W., Klin, A., Chawarska, K.: Pivotal response treatment for infants at-risk for autism spectrum disorders: a pilot study. J. Autism Dev. Disord. **43**(1), 91–102 (2013)
118. Suhrheinrich, J., Chan, J.: Exploring the effect of immediate video feedback on coaching. J. Spec. Educ. Technol. **32**(1), 47–53 (2017)
119. Suhrheinrich, J., Reed, S., Schreibman, L., Bolduc, C.: Classroom pivotal response teaching for children with autism. Guilford Press, New York (2011)
120. Symon, J.B.: Expanding interventions for children with autism: parents as trainers. J. Positive Behav. Interv. **7**(3), 159–173 (2005)
121. Tamura, Y., Yano, S., Osumi, H.: Modeling of human attention based on analysis of magic. In: Proceedings of the 2014 ACM/IEEE international conference on Human–robot interaction, pp. 302–303. ACM (2014)
122. Tamura, Y., Akashi, T., Yano, S., Osumi, H.: Human visual attention model based on analysis of magic for smooth human–robot interaction. Int. J. Soc. Robot. **8**(5), 685–694 (2016)
123. The Help Group: Parenting classes. http://www.thehelpgroup.org/parent/parenting-classes/. Accessed 19 May 2018 (2018)
124. Thorp, D.M., Stahmer, A.C., Schreibman, L.: Effects of sociodramatic play training on children with autism. J. Autism Dev. Disord. **25**(3), 265–282 (1995)

125. Tripathi, S., Lipton, Z.C., Belongie, S., Nguyen, T.: Context matters: Refining object detection in video with recurrent neural networks. Preprint arXiv:1607.04648 (2016)
126. Tsatsoulis, P.D., Kordas, P., Marshall, M., Forsyth, D., Rozga, A.: The static multimodal dyadic behavior dataset for engagement prediction. In: Computer Vision-ECCV 2016 Workshops, pp. 386–399. Springer (2016)
127. University of Washington: Parent family trainings. https://depts.washington.edu/uwautism/training/uwac-workshops/parentfamily/. Accessed 19 May 2018 (2018)
128. Van Gemeren, C., Poppe, R., Veltkamp, R.C.: Spatio-temporal detection of fine-grained dyadic human interactions. In: International Workshop on Human Behavior Understanding, pp. 116–133. Springer (2016)
129. Ventola, P., Friedman, H.E., Anderson, L.C., Wolf, J.M., Oosting, D., Foss-Feig, J., McDonald, N., Volkmar, F., Pelphrey, K.A.: Improvements in social and adaptive functioning following short-duration PRT program: a clinical replication. J. Autism Dev. Disord. **44**(11), 2862–2870 (2014)
130. Vismara, L.A., Lyons, G.L.: Using perseverative interests to elicit joint attention behaviors in young children with autism: theoretical and clinical implications for understanding motivation. J. Positive Behav. Interv. **9**(4), 214–228 (2007)
131. Vismara, L.A., Young, G.S., Stahmer, A.C., Griffith, E.M., Rogers, S.J.: Dissemination of evidence-based practice: Can we train therapists from a distance? J. Autism Dev. Disord. **39**(12), 1636 (2009)
132. Vismara, L.A., Young, G.S., Rogers, S.J.: Telehealth for expanding the reach of early autism training to parents. Autism Res. Treat. **2012**, 12 (2012)
133. Vismara, L.A., McCormick, C., Young, G.S., Nadhan, A., Monlux, K.: Preliminary findings of a telehealth approach to parent training in autism. J. Autism Dev. Disord. **43**(12), 2953–2969 (2013)
134. Wang, Y., Berant, J., Liang, P.: Building a semantic parser overnight. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 1: Long Papers), pp. 1332–1342 (2015)
135. Wei, P., Xie, D., Zheng, N., Zhu, S.C.: Inferring human attention by learning latent intentions. In: Proceedings of the Twenty-Sixth International Joint Conference of Artificial Intelligence (2017)
136. Wong, C., Odom, S.L., Hume, K.A., Cox, A.W., Fettig, A., Kucharczyk, S., Brock, M.E., Plavnick, J.B., Fleury, V.P., Schultz, T.R.: Evidence-based practices for children, youth, and young adults with autism spectrum disorder: a comprehensive review. J. Autism Dev. Disord. **45**(7), 1951–1966 (2015)
137. Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., Hansen, J.: Signal processing for young child speech language development. In: First Workshop on Child, Computer and Interaction (2008)
138. Xu, D., Gilkerson, J., Richards, J., Yapanel, U., Gray, S.: Child vocalization composition as discriminant information for automatic autism detection. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pp. 2518–2522. IEEE (2009)
139. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. Preprint arXiv:1708.02709 (2017)
140. Zhang, X.L., Wang, D.: Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **24**(2), 252–264 (2016)
141. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: European Conference on Computer Vision, pp. 707–721. Springer (2012)
142. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y., Courville, A.: Towards end-to-end speech recognition with deep convolutional neural networks. Preprint arXiv:1701.02720 (2017)
143. Zhang, Y., Yan, D., Yuan, Y.: An object tracking algorithm with embedded gyro information. In: Seventh International Conference on Electronics and Information Engineering, vol. 10322, p. 103220U. International Society for Optics and Photonics (2017)
144. Zhao, R., Ali, H., van der Smagt, P.: Two-stream RNN/CNN for action recognition in 3d videos. Preprint arXiv:1703.09783 (2017)