# Multimodal Data Analysis of Dyadic Interactions for an Automated Feedback System Supporting Parent Implementation of Pivotal Response Treatment

by

Corey D Copenhaver Heath

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2019 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
Troy McDaniel
Hemanth Venkateswara
Hasan Davulcu
Ashraf Gaffar

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Parents fulfill a pivotal role in early childhood development of social and communication skills. In children with autism, the development of these skills can be delayed. Applied behavioral analysis (ABA) techniques have been created to aid in skill acquisition. Among these, pivotal response treatment (PRT) has been empirically shown to foster improvements. Research into PRT implementation has also shown that parents can be trained to be effective interventionists for their children. The current difficulty in PRT training is how to disseminate training to parents who need it, and how to support and motivate practitioners after training.

Evaluation of the parents' fidelity to implementation is often undertaken using video probes that depict the dyadic interaction occurring between the parent and the child during PRT sessions. These videos are time consuming for clinicians to process, and often result in only minimal feedback for the parents. Current trends in technology could be utilized to alleviate the manual cost of extracting data from the videos, affording greater opportunities for providing clinician created feedback as well as automated assessments.

The naturalistic context of the video probes along with the dependence on ubiquitous recording devices creates a difficult scenario for classification tasks. The domain of the PRT video probes can be expected to have high levels of both aleatory and epistemic uncertainty. Addressing these challenges requires examination of the multimodal data along with implementation and evaluation of classification algorithms. This is explored through the use of a new dataset of PRT videos.

The relationship between the parent and the clinician is important. The clinician can provide support and help build self-efficacy in addition to providing knowledge and modeling of treatment procedures. Facilitating this relationship along with automated feedback not only provides the opportunity to present expert feedback to the parent, but

also allows the clinician to aid in personalizing the classification models. By utilizing a human-in-the-loop framework, clinicians can aid in addressing the uncertainty in the classification models by providing additional labeled samples. This will allow the system to improve classification and provides a person-centered approach to extracting multimodal data from PRT video probes.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

# Chapter 1

# INTRODUCTION

In early childhood development, the people with the most regular interactions with a 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.

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 help children improve social and communication skills, 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 exacerbated as diagnoses of ASD continue to increase.

Parents and caregivers are able to quickly learn to implement ABA techniques effectively; however, fidelity to the methodologies tends to drop shortly after the training period. This could be due to a number of factors, including a lack of practice, failure to adapt the procedure, or the parent having insufficient confidence in his or her ability to produce the desired result. Ideally, this would be addressed by continual support from the

clinician, but maintaining this connection is subject to logistical constraints. The most common constraints are the time-cost from the clinician to provide evaluation and feedback, and the limited resources available, particularly in rural communities. In particular, evaluating implementation fidelity relies on manually extracting performance metrics from video probes of the caregiver and child. To scale the system to support more individuals, and make maintaining relationships through remote connections easier, it is important to look at the ways technology can be incorporated into the system to ensure people have access to training materials and implementation feedback.

Preserving the relationship between the parent and the clinician is an important part of the process. This relationship capitalizes on the expertise of the clinician and the human connection to add social pressure to continue utilizing treatments. Using a human-in-the-loop design paradigm affords the opportunity of using technology to reduce the human costs of evaluating and supporting individuals learning ABA methodologies while also facilitating a meaningful connection with the clinical experts.

This dissertation explores creating a human-in-the-loop feedback system for supporting caregivers. This work focuses on the clinician perspective of the project by examining the metrics and tools that can be utilized to reduce the manual cost of providing feedback. In addition to reducing the human cost of evaluation, using automated techniques allows additional performance data to be recorded that could provide new insights into the interactions. These metrics could be used in aiding the clinician in analysis and automatically tracking the progress of the child's communications skills as he or she improves. This information would be important for determining when learning plateaus are reached, requiring a change in the skills being targeted.

To approach this system, the following overarching research questions were proposed:

- What technologies and approaches are currently used to train caregivers in ABA methodologies?
- How can current evaluation metrics for PRT fidelity be collected automatically using current trends in artificial intelligence research?
- What additional metrics could be extracted from the videos that would aid the clinician in providing feedback?
- How can a human-in-the-loop system be designed to facilitate the connection between the clinician and caregivers?
- What affordances does the continuous involvement of the clinician with the automated system provide in terms of personalizing classification?

In exploring these questions, it became apparent that the PRT video probes presented challenges that were not being adequately addressed in the current literature. The video data presented adverse recording conditions, requiring video processing and vision-related classification techniques to be robust to camera movement, occlusion, and a diverse set of activities. Similarly, audio tracks could be subject to variable recording conditions, often making it difficult to discern child vocalizations. In addition to this, PRT sessions depict speaking patterns not common in conversational speech. Caregivers often use child-directed speech, or baby talk, to engage the child. It is also important to detect not only when a child speaks a full word, but any attempts at vocalization. Before undertaking work on a feedback system, these classification tasks needed to be addressed.

The research presented focuses on one common performance metric utilized in fidelity evaluations in research studies and training courses - creating an 'opportunity to respond.' Determining if the caregiver has correctly demonstrated creating an

‘opportunity to respond’ requires detecting the attention state of the child and analyzing the caregiver’s instruction.

## 1.1 Contributions

Pursuing this work has resulted in nine primary contributions to computer science. First, a comprehensive review of the literature, including technologies for ABA training, computer vision techniques for detecting dyadic interactions and joint attention, the application of computer vision techniques to videos depicting children with autism, and audio processing techniques for voice activity detection, speaker separation, and child speech recognition.

Second, a new dataset consisting of video probes from PRT sessions between parents and their children was created. These videos represent an ‘in-the-wild’ scenario for classification and recognition tasks. The dataset was labeled for the child’s attention state based on the visual data. The audio was transcribed and annotated based on the speaker. This transcription included nonspeech vocalizations made by the child.

Third, the visual frames from the video were analyzed to determine a method for detecting the child’s attention state during his interaction with his parent. This was used to evaluate classifying attention in video data using spatial and contrived features as input for machine learning algorithms.This process provides a baseline for evaluating approaches to detecting attention in ABA interactions, and could be applied to similar domains.

Fourth, a novel voice activity detection and speaker separation scenario was explored. The audio from the PRT probes exhibits atypical speech patterns that are not considered in research focusing on automated speech recognition, including child-directed speech patterns and non-speech vocalizations.

Fifth, the use of multimodal data for detecting attention and inferring when an ‘opportunity to respond’ has occurred was explored. The contributions of this aspect of

the project examined feature fusion techniques and the use of confidence estimates of class predictions for classification tasks involving audio and video data.

Sixth, evaluation of automatically gathered metrics and a user interface for clinicians to provide feedback was developed in collaboration with behavior analysts. This provided insight into how the system could be effectively used. No similar research projects have been encountered in the literature.

Seventh, the continuous interaction of both the clinician and the parent in the proposed feedback system provides the opportunity to use the expertise of the clinician to personalize classification models for the parent. This is a novel aspect of the system and has been explored for its efficiency and feasibility.

Eighth, abbreviating the videos is important for reducing the processing time for clinicians. An approach is introduced that uses a graph-based representation of the individuals to identify keyframes. This is compared to color histogram approaches to examine robustness to camera instability and occlusion. This is a novel application that could be beneficial to other applied domains.

Ninth, the conclusion of this work provides a discussion of other approaches for classification, congruent applications, and additional ways to incorporate technology into PRT.

## 1.2 Dissertation Overview

**Chapter 2** provides an overview of relevant literature for the project and how technology could be applied to automate data collection of performance metrics that are currently being extracted manually. This chapter discusses both contrived and naturalistic ABA methodologies and presents the current literature regarding the use of technology for ABA training. Using video probes is an important method for discerning fidelity to ABA techniques. The data collection methodologies and performance evaluation criteria are

presented. The chapter concludes by introducing concepts from computer science research that could be utilized to automate data collection on each critium.

**Chapter 3** explores the application of computer vision and video data processing to the PRT videos. This chapter includes relevant research on image and video processing and classification that could be utilized in an automated feedback system for PRT. Of primary interest to the project are techniques for classifying dyadic actions and detecting joint attention. The PRT video dataset is described and analyzed based on extracted features from the individuals depicted in each frame. A comparison of techniques for classifying child attention are presented and evaluated.

**Chapter 4** examines the application of machine learning techniques for voice activity detection and speaker separation. A review of relevant literature is presented regarding automated speech recognition techniques that could be utilized to evaluate caregiver instructions and child responses. Multiple techniques for extracting the adult and child vocalizations are evaluated to determine an effective method for processing the PRT audio data.

**Chapter 5** presents research regarding multimodal fusion techniques and evaluating prediction confidence estimates. Determining if an opportunity to respond has occurred is a multimodal task, depending on both the audio and video data. This also provides the opportunity to explore how the audio data could be used to improve attention classification. Different feature selection techniques, fusion methods, and sample sizes are examined and compared to determine the most effective method for multimodal classification tasks.

**Chapter 6** extends the research regarding the role of the clinician in the human-in-the-loop system. The main research objective is to examine methods for personalizing classification to the parent-child dyad. It was surmised that the clinician, as an expert on the interactions, can provide additional information to the system to improve

the performance on attention detection. This chapter explores the feasibility of different approaches in order to balance the effort of the clinician with the benefits to classification.

**Chapter 7** introduces a prototype user interface (UI) for the clinicians to examine the initial steps toward automated assessment of the video probes. The UI prototype was designed using agile methodologies and incorporated feedback from the clinicians through the development process in conjunction with participatory design. The goal was to use the development of the prototype to facilitate discovering the clinician's wants and needs for the system, as well as the benefits of the automatically collected data. This helped identify meaningful data visuals and video segment sizes. The clinicians expressed enthusiasm toward the system.

**Chapter 8**, as the final chapter, presents additional ideas that were outside of the scope of the dissertation but important to the problem. This consists of presenting how the system could be expanded, different approaches to classification, and parallels to other domains. This chapter also discusses how the work presented in this dissertation fits into a feedback system designed to facilitate self-regulatory learning.

## 1.3 Previously Published Work

The contents of this dissertation have been partially published in conferences and journal publications. Information presented in Chapter 2, along with background research for Chapters 3 and 4, was published in *"Improving Communication Skills of Children With Autism through Support of Applied Behavior Analysis Treatments using Multimedia Computing: A Survey"* (Heath et al., 2019b). The experiments and results from Chapter 3 were published in *"Detecting Attention in Pivotal Response Treatment Video Probes"* (Heath et al., 2018). Chapter 4 was published in *"Parent and Child Voice Activity Detection in Pivotal Response Treatment Video Probes"* (Heath et al., 2019c). Experiments regarding detecting opportunities to respond in Chapter 5 were published in

*"Using Multimodal Data for Automated Fidelity Evaluation in Pivotal Response Treatment Videos"* (Heath et al., 2019d). The user interface design and evaluation presented in Chapter 6 were published in *"Using Participatory Design to Create a User Interface for Analyzing Pivotal Response Treatment Video Probes"* (Heath et al., 2019a). Discussion of low power computing applications in PRT videos from Chapter 8 were published in *"Are You Paying Attention? Classifying Attention in Pivotal Response Treatment Videos"* (Heath et al., 2019e).

# Chapter 2

# TECHNOLOGIES FOR APPLIED BEHAVIOR ANALYSIS TRAINING AND IMPLEMENTATION

Applied Behavior Analysis (ABA) is an approach to creating and implementing procedures to promote beneficial behaviors and diminish disadvantageous behaviors. ABA focuses on applying a scientific methodology to behavior treatments, emphasizing replicable techniques, and the 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, which is its ability to be generalized to target different behaviors under differing circumstances (Baer et al., 1968).

There are two general approaches to designing ABA treatments - contrived and naturalistic (Kane et al., 2010). 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.

The intent of this chapter 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. Additionally, a comparison with DTT will be presented. Following the discussion on implementation, research regarding training non-clinicians, caregivers, teachers, peers, and paraprofessionals will be examined, including studies incorporating technology. The research presented forms a foundation for how technology could be utilized in PRT training and implementation.

## 2.1 Naturalistic Applied Behavior Analysis Implementation

Often described as a way of life rather than a teaching methodology (McClelland et al., 2016), naturalistic ABA is intended to be integrated with daily activities. These methods focus on keeping the recipient of the treatment active and making the learning activities relevant (Schreibman et al., 2015). The interactions in naturalistic ABA are undertaken

between an interventionist and a recipient. The interventionist engages with the recipient in an activity of the recipient's choice. Allowing the recipient to choose the activity makes it so that the technique capitalizes on the recipient's natural motivation to continue with that activity. By following the lead of the recipient, the interventionist presents learning opportunities based on the skills being targeted.

Implementation is based on a generalized three-part sequence of antecedent, behavior, and consequence. The antecedent focuses on the actions the treatment interventionist takes to prompt the recipient with a learning activity. First, this means gaining the recipient's attention, which generally involves seizing control of the object or activity the recipient is currently participating in. After gaining the recipient's attention, the interventionist can then give an instruction. Verbal instructions can include modeling the word or phrase the recipient is expected to say, saying the beginning of a sequence, such as counting, and 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 (Koegel, 1988; Koegel et al., 1988). Verbal instructions should be limited to the speaking level of the child.

The behavior is the recipient's reaction to the antecedent. Ideally the recipient 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 recipient's current skill level. If the recipient is mostly non-verbal, a correct response could be an attempt at speech or vocalizing the phoneme of the expected word. If the recipient 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 recipient 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 recipient completing an adequate attempt for his or her current skill level.

Outside of the antecedent, behavior, and consequence sequence, recipients should also be rewarded for initiating social interactions, asking questions, and spontaneous speech (Koegel et al., 2014b; Schreibman et al., 2015). Children with ASD can exhibit a deficit in initiating social interaction, so part of the naturalistic intervention should include creating situations that necessitate the recipient taking the initiative. This can include placing a favorite object, such as a toy, in a visible but unreachable location to encourage him or her 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 recipient in order to increase his or her ability. Skills that have been achieved by the recipient can become maintenance skills. Maintenance skills are intermixed with target skills during treatment sessions to ensure that the recipient continues to practice and to keep the recipient motivated by giving them a challenge they can overcome.

There are nearly 30 years of published research on naturalistic ABA, primarily PRT, mostly focusing on children with autism between the ages of 6 - 11 (Wong et al., 2015). 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 of 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 in numerous

publications (Koegel et al., 1997, 2003, 2010, 1999a,b, 2014b, 2009; Sherer and Schreibman, 2005; Ventola et al., 2014). Improvement in joint attention after naturalistic ABA intervention was reported by Vismara and Lyons (2007). Increases in social and symbolic play were published by Thorp et al. (1995) and Stahmer (1995); Stahmer et al. (2006). Studies examining the affective state of children also showed positive results following treatments. PRT was correlated with a reduction of anxiety in children by Lecavalier et al. (2017), resulting in less disruptive behaviors.

Studies conducted by Koegel et al. (1998, 1987) and Mohammadzaheri et al. (2014) compared PRT to DTT or a similar contrived ABA implementation to evaluate children's post-intervention communication skills. Looking at the mean number of spontaneous utterances, intelligibility, and mean length of vocal utterance, respectively, each study concluded that children who received PRT showed greater improvement than children who received a contrived ABA implementation. Similar conclusions were drawn regarding reduction of disruptive behaviors by Koegel et al. (1992) and Mohammadzaheri et al. (2015) 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 (Koegel et al., 1996; Schreibman et al., 1991) that concluded PRT was related to greater increases in happiness and reduction of stress of both parents and children compared to DTT interventions.

## 2.2 Comparison of Naturalistic and Contrived Techniques

Naturalistic and contrived techniques follow a similar structure. Looking at DTT as an example of a contrived technique, the overall methodology followers similar steps to PRT. DTT consists of an antecedent delivered by the interventionist that consists of two parts. The first part of the antecedent is the cue, or instruction. This is the action from the interventionist that is meant to elicit the desired behavior from the recipient. The

instruction needs to be clearly articulated and delivered when the recipient is attending the interventionist. The second component of the antecedent is a model prompt. This is delivered in a more structured manner in DTT than in PRT. In DTT, the prompt follows the instruction after new tasks are introduced, then gradually the prompt is faded in preceding intervals until the recipient is demonstrating the desired behavior on his or her own (Smith, 2001).

As in naturalistic methodologies, the antecedent is followed by the behavior, consisting of either a sufficient attempt at the desired skill or an incorrect response. After the response, the interventionist is expected to provide a consequence for the behavior. This should be delivered within three to five seconds (Sarokoff and Sturmey, 2004). The type of acceptable reward for correctly demonstrated behaviors can be anything the recipient finds pleasing or motivating, such as verbal praise, food rewards, or play time with preferred objects or toys. This is a fundamental difference with naturalistic techniques, where the recipient identifies and engages in a motivating activity on his or her own while the interventionist incorporates learning opportunities. Although broadening the types of rewards that can be implemented gives interventionists more freedom, it does require additional effort to discover rewards that will provide adequate motivation for the recipient (Leaf et al., 2015).

An additional formal component of DTT that is less defined in naturalistic methodologies is the iteration interval. DTT is drill-based. The interventionist reiterates exercises in succession in order to facilitate learning by repetition. The exercises, or trials, are intended to be short to allow for maximum repetition. A short rest period of one to five seconds is recommended between iterations (Smith, 2001).

The design of DTT is meant to provide individualized and simplified instruction presented in a one-on-one environment (Smith, 2001). The structured implementation provides three advantages compared to naturalistic methodologies. First, the intervention

session, along with the trials themselves, have a clear start and stop time. This helps motivate recipients by reinforcing that the tasks will be limited in duration. Second, following a formal procedure makes data collection easier (McClelland et al., 2016). Having the distinct trials allows for more structure in collecting metrics based on the recipient's performance on the given tasks. With a clear start and stop point, individuals evaluating interventionist fidelity can also clearly discern when actions are taking place. Third, the interventionist selects the activities for the intervention, allowing him or her to create lesson plans and objectives. Conversely, PRT is based on activities selected by the child, requiring the interventionist to improvise learning opportunities when engaging in new tasks.

The primary drawback of contrived techniques is that it removes learning from its context. By creating drill-based learning regimes, the recipient may be able to perform exceptionally in the learning context but be unable to generalize the concepts and transfer the learning to related tasks (Smith, 2001).

Studies have shown that DTT is effective at training recipients, particularly school-aged children, in a variety of tasks. These include social and play skills (Lovaas and Smith, 2003; Smith, 2010), receptive discrimination and academic skills (Gutierrez Jr et al., 2009; Sarokoff and Sturmey, 2008; Skokut et al., 2008; Sturmey and Fitzer, 2007), vocal response (Young et al., 1994), play activities (Coe et al., 1990; Lovaas and Smith, 2003; Smith, 2010), and reduction of self-harming behavior (Matson and LoVullo, 2008). This differs from much of the naturalistic literature which is dominated by research into improving communication and social skills.

Neither contrived nor naturalistic methodologies are inclusive, and it is often useful to employ multiple strategies to cover various skills and contexts. In particular, DTT is useful for teaching the recipient to sit still and listen to the instructor (McClelland et al., 2016). These skills would be more difficult to teach under a naturalistic setting. PRT has

been favored for caregiver implementation because of the emphasis on child motivation and the opportunity caregivers have to imbed learning opportunities in play activities. All intervention techniques require a significant amount of time per week to be effective. Making the activity more enjoyable will aid in more frequent implementation.

## 2.3 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 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. (Baker-Ericzén et al., 2007; Coolican et al., 2010; Gillett and LeBlanc, 2007; Hardan et al., 2015; Laski et al., 1988; Smith et al., 2015). The improvements associated with PRT training for parents was concluded to be independent of age, gender, or ethnicity (Baker-Ericzén et al., 2007). Attempts to start interventions early in the child's development has also fueled research into training parents to implement interventions. Positive effects for infants after parents received naturalistic ABA training were reported by Steiner et al. (2013) and Koegel et al. (2014a).

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-aged children with ASD (Jones and Feeley, 2009; Kasari et al., 2015). Teacher implementation of JASPER with preschool-aged children with ASD also showed positive effects on joint attention, with noteworthy effects on child-initiated joint attention (Lawton and Kasari, 2012).

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 was noted after training in PRT (Steiner et al., 2013) and ESDM (Estes et al., 2014). This is particularly important since parents of children with ASD report high levels of stress (Johnson et al., 2011; Kasari et al., 2015; Nefdt et al., 2010). 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 (Kasari et al., 2015).

Peer implementation of PRT for elementary school students has also been explored (Harper et al., 2008; Pierce and Schreibman, 1995, 1997). 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 recipient generalize social skills (Pierce and Schreibman, 1997). 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 (Harper et al., 2008).

## 2.4 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 2.1 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 hours, with the most common duration being 12 hours. The study from Jones and Feeley (2009) was not included in these calculations. The participants in the study received one hour 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.

Table 2.1
Publications on training non-clinicians in ABA implementation

| Publication | Training Method | Participants | Training Duration (hours) |
|---|---|---|---|
| Laski et al. (1988) | In-person Clinician Instruction | Parents | 3.75[a] |
| Pierce and Schreibman (1995) | In-person Clinician Instruction | Peers | 2 |
| Pierce and Schreibman (1997) | In-person Clinician Instruction | Peers | 2 |

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 four, six, and 12 hours of training respectively

Table 2.1 Coninuted: Publications on training non-clinicians in ABA implementation

| Publication | Training Method | Participants | Training Duration (hours) |
|---|---|---|---|
| Koegel et al. (2002) | In-person Clinician Instruction | Parents | 25 |
| Symon (2005) | In-person Peer Instruction | Parents | 25 |
| Baker-Ericzén et al. (2007) | In-person Clinician Instruction | Parents | 12 |
| Gillett and LeBlanc (2007) | In-person Clinician Instruction | Parents | 3 |
| Harper et al. (2008) | In-person Clinician Instruction | Peers | 1 |
| Jones and Feeley (2009) | In-person Clinician Instruction | Parents | 1[b] |
| Vismara et al. (2009) | Tele-conference / Video Instruction | Therapists | 17 |
| Coolican et al. (2010) | In-person Clinician Instruction | Parents | 6 |
| Machalicek et al. (2010) | Tele-conference | Teachers | 1.25[c] |
| Nefdt et al. (2010) | Video Instruction | Parents | 1.6 |
| Lawton and Kasari (2012) | In-person Clinician Instruction | Teachers | 6 |
| Vismara et al. (2012) | Tele-conference | Parents | 12 |
| Vismara et al. (2013) | Tele-conference / Video Instruction | Parents | 12 |
| Steiner et al. (2013) | In-person Clinician Instruction | Parents | 10 |
| Estes et al. (2014) | In-person Clinician Instruction | Parents | 12 |
| Koegel et al. (2014a) | In-person Clinician Instruction | Parents | 7.33[d] |
| Kasari et al. (2015) | In-person Clinician Instruction | Parents | 10 |
| Gengoux et al. (2015); Hardan et al. (2015) | In-person Clinician Instruction | Parents | 16 |

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 four, six, and 12 hours of training respectively

Table 2.1 Coninuted: Publications on training non-clinicians in ABA implementation

| Publication | Training Method | Participants | Training Duration (hours) |
|---|---|---|---|
| Smith et al. (2015) | In-person Clinician Instruction | Parents | 8 |
| Suhrheinrich and Chan (2017) | 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 four, six, and 12 hours of training respectively

The duration metrics from Table 2.1 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 United States shows training options are typically centered around group workshops or one-on-one support sessions. For example, an eight hour group course teaching ABA is offered by the University of Washington Autism Center[1] in Seattle. Twenty hour group courses are offered by The Help Group[2] in Sherman Oaks, California and the Southwest Autism Resource and Research Center (SARRC)[3] in Phoenix, Arizona on skills for parenting children with developmental disabilities and PRT implementation, respectively. The Children's Hospital at Sacred-Heart[4] offers a 12 hour ABA implementation course in Pensacola, Florida.

[1]https://depts.washington.edu/uwautism/training/uwac-workshops/parentfamily/
[2]http://www.thehelpgroup.org/parent/parenting-classes/
[3]https://autismcenter.org/parents-and-caregivers
[4]https://sacred-heart.org/childrenshospital/main/services/

One-on-one training courses were advertised by SARRC and the Choice Autism Center[1] in Traverse City, Michigan. The SARRC website listed an individualized 12 session, one hour 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 five years, a 20 to 40 hour per week program is listed. A six to 20 hour per week program is available for children ages six 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 with medical centers, universities, or research institutions, often located in larger metropolitan areas. Individuals in rural or remote

[1] https://choiceautismcenter.com/our-programs/

locations may not have a resource center in the immediate area and may not be able to travel to receive training.

## 2.5 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 Symon (2005) 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. (2009) examined technology for remote instructions for training therapists to teach parents to implement ESDM (Vismara et al., 2009). 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. (2012) applied this telecast training model to teach parents 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. (Vismara et al., 2013) 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. (2010). 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 to 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 setup 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 (Bagaiolo et al., 2017). 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%-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. (2010) 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 (Gillesen et al., 2011). 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.

## 2.6 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% (Koegel et al., 2002). 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. Although these categories are often adapted to fit the intervention methodology and the child skill being targeted, they typically consist of the following: delivery of a clear instruction, diversity of tasks, following the recipient, identifying natural motivators, contingency, and reinforcing attempts (Koegel et al., 2002; Nefdt et al., 2010).

Delivering a clear instruction 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 (Koegel, 1988; Leaf et al., 2016; Suhrheinrich et al., 2011). 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 delivering a clear instruction is the instruction itself. 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 diversity of tasks, 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 it 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, whereas a recipient with more verbal skills would need to speak the full word or phrase for it 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 (Suhrheinrich and Chan, 2017).

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.

## 2.7 New Directions for Incorporating Technology

The research presented above illustrates some of the challenges faced by behavior 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 has been scrutinized. To address location constraints, researchers have designed condensed courses (Koegel et al., 2002), and implemented live teleconferencing (Vismara et al., 2012). 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 longer. What all of these approaches lack is long-term feedback. The duration of training programs often last only 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 goals; however, they may be unable to 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 greater 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 regards 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 (Vismara et al., 2013) 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 2.2 provides a brief overview of the areas of multimedia processing that could be utilized to extract information from video probes in regards to the current human evaluation-based scoring methodologies. These scoring methodologies are multimodal and depend on both visual and auditory 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.

Table 2.2
Naturalistic ABA evaluation criteria and relevant areas of technology that could be applied for automated analysis.

| Evaluation Category | Category | Relevant Areas of Technology | How It Could Be Implemented | Feasibility (High,Medium,Low) |
|---|---|---|---|---|
| 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 vs. 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 Reinforce-ment | 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 |

## 2.8 Current Feasibility of Automating Evaluation Assessments

The evaluation criteria presented in Table 2.2 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 versus target skills. 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 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 if there is sufficient variation in the instructions. Evaluating if 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 non-speech 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 if the reinforcement is the continuation of an activity, or if 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 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 criteria 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 it, the interaction may not be difficult to classify. In this instance, attention can be inferred by determining if 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.

## 2.9 Current Work

The objective of the dissertation is to detect when an 'opportunity to respond' has occurred. To accomplish this, the video and audio data needs to be analyzed, and classification models need to be trained. The most pertinent tasks are to detect the child's attention to the parent and extract the parent's instructions from the audio track. Chapter 3 will discuss analyzing the video data to classify the child's attention state. Chapter 4 will examine the audio data to perform voice activity detection and speaker separation. Chapter 5 will present how both video and audio data channels can be utilized to improve attention classification and make an inference on when the parent has provided a proper 'opportunity to respond.'

# Chapter 3

# DETECTING CHILD ATTENTION IN PRT VIDEO PROBES

Video probes present a challenge for automated analysis. Ideally, the videos should show the interventionist (the caregiver), and the recipient (the child), completely in frame, unobscured, and facing the camera. This is not always the case. These videos are often filmed using handheld digital cameras or mobile phone cameras. The resulting videos are often low resolution, unstable, and occasionally have the interventionist or the recipient out of the frame. The video probes are often recorded at home in an unstructured environment with inconsistent backgrounds, which could provide a challenge for computer vision-based algorithms (Brutzer et al., 2011). Additionally, because PRT is based on integrating learning opportunities within activities selected by the treatment recipient, there are no standardized actions or activities reflected in the videos.

This chapter explores how PRT performance data can be extracted from the video probes automatically to reduce the processing time for clinicians and provide feedback to PRT interventionists. The research presented focuses on one evaluation metric utilized for feedback - gaining the recipient's attention. This is an important step in training the interventionist to provide proper instruction.

For the analysis, a new labeled dataset consisting of body pose data from PRT video probes was created. Strategies were examined for preparing data and approximating data gaps in natural, untrimmed videos, along with methods for building spatio-temporal (ST) graphs for dyadic interactions. Additionally, the implementation of a machine learning model for detecting attention is explored, through the comparison of Support Vector Machine (SVM) implementations using Euclidean-based data and a pretrained convolutional neural network (CNN) model with AlexNet weights fine-tuned with pixel data from video probe still frames. This implementation is based on research video

processes and computer vision. The models utilized in the presented approach are intended to serve as a baseline for future innovation. Ultimately, I present how automated detection of attention can be used to aid clinicians and caregivers reviewing PRT video probes.

## 3.1 Research in Video Processing

There are several opportunities to incorporate video processing techniques into the analysis of PRT videos. Relevant research regarding object tracking, human pose estimation, activity detection, and attention classification will be presented in the subsequent section. This research provides the background for analyzing the dataset and presenting preliminary classification models for assessing child attention in the video probes.

### *3.1.1 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/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 (Cao et al., 2017). 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 (Itti et al., 1998). In video, changes between frames adds 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 (Koh and Kim, 2017). It is then presumed that moving objects of the video are important and garner the viewer's attention (Tamura et al., 2016, 2014).

PRT videos are a challenging medium for applying object tracking. 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 could be a method in 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 if the parent is following the child's lead or providing the child the reinforcer as part of the consequent 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 attention, activities being performed, and the dyadic interaction between the parent and child.

### *3.1.2 Human Pose Estimation*

Pose estimation tasks focus on detecting the articulation of human figures in an image. The most commonly used approach identifies the human in the image, then uses the pictorial structure to create a graph representing key body points (Andriluka et al., 2009, 2010; Felzenszwalb and Huttenlocher, 2005; Pishchulin et al., 2013). In the approach, the body is divided into a set of regions representing individual parts, such as the arms and

legs. Detecting the parts is based on the assumption that the figures depicted will follow a general model of the human body. This allows for the discriminators to use inference to select body parts based on their proximity to other parts. This approach works best when an individual's full body is visible in the frame (Wei et al., 2016).

Learning the individual body parts separate from the full body pose can improve on pictorial models. This builds on the previous idea of creating a model for the human form by organizing the model based on the recognition difficulty. This allows the pose estimation algorithms to utilize the classification of larger body parts to find more obscure parts. This particularly attempts to address misclassifying the image background as body features (Sun and Savarese, 2011; Tian et al., 2012).

Deep learning methods have focused on using convolutional networks for identifying keypoint location, usually articulation points, in the images (Toshev and Szegedy, 2014; Wei et al., 2016). These approaches follow a sequential process for detecting the pose, often identifying a key articulation point and expanding outward. This was accomplished in (Toshev and Szegedy, 2014) by expanding the subsample size for subsequent convolutional layers for areas around a potential articulation point. Similar to this, Einfalt et al. (2018) used heat maps identified around key body points across multiple frames to infer point locations in videos of swimmers. The heatmaps allowed for the construction of pose graph candidates that could then be selected based on the algorithm's confidence levels.

Pose extraction for this project was undertaken using OpenPose. OpenPose uses a convolutional approach to pose estimation (Cao et al., 2017; Wei et al., 2016). The algorithm creates a map of the probability of a key body point being represented at a particular pixel location. After identifying these points, their association with adjacent detected points is inferred based on creating a vector extending in the likely direction of

the next distal anatomical feature. Multiple individuals in the frame are detected based on the detection of multiple instances of the same body points.

### *3.1.3 Human Activity Detection*

Building on the concepts of object tracking and human pose estimation, activity detection and classification in video data relies on analyzing both spatial information about the configuration of people and objects in an individual frame. Temporal information is relayed based on 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 (Raptis and Sigal, 2013), organizing frames into a graph for analysis (Chen and Grauman, 2017), examining sequences of identified images (Ma et al., 2016; Tripathi et al., 2016), observing object flow (Voulodimos et al., 2016), and exploring spatio-temporal representations of the individuals in a frame (Fragkiadaki et al., 2015; Jain et al., 2016). Research on applying activity classification to individuals with autism has focused on single person activities, primarily detecting self-stimulating or repetitive behaviors (Coronato et al., 2014; Jazouli et al., 2016; Khan et al., 2017).

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.

### *3.1.4 Dyadic Activity Detection*

Activity classification for two or more people can take two common approaches. The activities of each individual can be classified separately then used to deduce a label about the entire scene, or the individuals can be analyzed as a single unit. In the first approach, the individuals in the frame are identified separately and their actions, position, and orientation in relation to other individuals is used to infer or describe the interactions (Bazzani et al., 2013; Deng et al., 2016, 2015; Hoai and Zisserman, 2014).

An additional 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 data for training a classification model (Van Gemeren et al., 2016; Zhang et al., 2012).

### *3.1.5 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 (Baxter et al., 2015; Duffner and Garcia, 2016; Wei et al., 2017) and surmise social engagement (Bazzani et al., 2013; Sener and Ikizler-Cinbis, 2015). 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 (Pusiol et al., 2014). 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 (Foster et al., 2013; Li et al., 2012; Sanghvi et al., 2011) and follow a similar

methodology to classification techniques for human interactions. Sanghvi et al. (2011) examined the visual cues of a child's state of engagement with a robot (Table 3.1).

Table 3.1
Visual signs that a child was engaged with the robot in Sanghvi et al. (2011).

| Engaged | Visual Cue |
|---|---|
| Yes | Anything except leaning forward<br>Little or no body motion<br>Rocking with mean body angle upright or backward<br>Little or no movement<br>Upward, forward, or backward movement<br>Upright posture<br>Leaning back or upright with hands in lap |
| No | Leaning forward (focus on game board)<br>Continuous periods of body motion<br>Leaning forward, rocking with mean angle leaning forward<br>Hand movement, or hands up, away from lap<br>External distractions, interactions with other people |

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 (Presti et al., 2013; Tsatsoulis et al., 2016). In addition to this, rate of movement has also been determined to be diagnostic of attention (Rajagopalan et al., 2016, 2015). Intuitively, an individual attending to another individual would likely not be changing position or orientation rapidly.

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 is 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 (Heath et al., 2018). This is possible by leveraging the domain knowledge that the frame should depict two individuals.

## 3.2 Proposed Approach

This chapter primarily examines a PRT video dataset and explores how a machine learning model could be developed to detect the recipient's attention. A new dataset was created based on existing videos of parents employing PRT with their child. Currently, no other datasets are equivalent to this task. The multimodal dyadic behavior (MMDB) dataset (Rehg et al., 2014) is the most similar to the conditions being examined in this project. The MMDB dataset depicts a child with ASD, sitting with his or her parent, interacting with a clinical professional in a limited set of activities. The dataset is labeled based on the child's visual display of engagement. The dataset exhibits a controlled environment with the data collection occurring under laboratory conditions. The interactions between the child and the clinician were filmed using multiple stationary cameras and an egocentric camera worn by the clinician. The participants were seated throughout the interactions. This dataset was used in several of the aforementioned research on engagement and joint attention (Rajagopalan et al., 2016, 2015; Tsatsoulis et al., 2016).

PRT sessions are expected to occur under ‘in-the-wild’ conditions. This means that the activities cannot be determined prior to the session, and the recording quality cannot be guaranteed. This differs from the MMDB conditions which controlled for activities and recording quality. In order to better represent the problem of classifying attention in PRT videos, a dataset consisting of labeled video segments of PRT sessions was created and a feature set of body and facial landmark points was extracted for each individual in the videos. This feature set was then examined for completeness and a post-processing procedure was run to approximate missing information. The augmented feature set was then used to train a nu-SVM model to classify video sequences based on the attention of the recipient. The SVM model was evaluated using a leave-one-out strategy where two videos were reserved as a validation set for each training iteration. The remaining sections describe the dataset and evaluations methodologies that were employed.

### *3.2.1 Dataset*

Fourteen videos were selected at random from a PRT study (Signh, 2014) to create a dataset for detecting a child’s attention to their caregiver. Each of the seven caregiver-child dyads were in two videos - a baseline video recorded before the caregiver received instruction in PRT and a post-study video recorded at the end of the training. In order to represent the types of videos that could be expected, no regard was paid to quality or levels of occlusion. The videos depict the child and caregiver engaged in various activities including playing with assorted toys and games, spinning in a chair, moving about the room, and watching videos on a cell phone. Each video was recorded by a clinician in a room at an ASD resource center.

The videos were divided into 30 frame increments for labeling. Labels were assigned according to the attentive state of the child in accordance with PRT literature (Koegel, 1988; Suhrheinrich et al., 2011). The segments were labeled as ‘attentive’ if the caregiver

controlled the motivator, and the child was not engaged in an activity, was looking at the parent, or was reaching toward the caregiver at any time during the video clip. For example, this is demonstrated in Figure 3.1 when the interventionist, the woman on the left, presents a motivating object to the recipient, the woman on the right. The segment was labeled 'inattentive' if the child controlled the motivator, was moving about the room, or was engaged in a solitary activity such that if the caregiver seized control of the motivator it would disrupt the activity. Figure 3.2 exemplifies inattention, showing the recipient has control over the motivational object, the puppet. Segments were labeled as 'shared attention' if the caregiver and child were engaged in a joint activity, or if the caregiver had control of a motivator in a way that seizing control was not disruptive to the activity. Shared attention is shown in Figure 3.3 when both the recipient and the interventionist have their own puppets and are participating in a mutual play activity. This allows the interventionist to present learning opportunities in the context of the play activity, with minimal disruption. Segments were ignored if either the child or caregiver were not visible.

The videos were processed using OpenPose (Cao et al., 2017) to extract spatial information to use for classification. OpenPose provides Cartesian data points for body and face landmarks for each individual identified in the frame. For the individual's body, 18 points are detected including the eyes, nose, neck, and major limb joints. Seventy additional points from detected facial landmarks are also recorded. The image on the left of Figure 3.4 depicts a frame from a PRT implementation video (Considine, 2011) where the individuals have the OpenPose feature graph overlaid.

To overtly capture the interaction between the two individuals in each frame, the Euclidean distance between the individuals' hands was calculated and provided as an additional feature for classification. The goal was to capture common motions that could

Figure 3.1: Screenshot from a PRT training video. The recipient (right) is in an attentive state, as indicated by her looking directly at the interventionist (left). Image from (Virgir05, 2015)

be indicative of attention, such as the child reaching toward an object in the caregiver's hand or the child playing with an object on his or her own.

Visual attention is an important feature for determining an individual's focus. An individual's gaze was estimated by calculating the Euler angles of the face using 15 of the facial features identified by OpenPose and an approximated camera perspective based on frame dimensions. The pitch and yaw angles were used to create a point projected away from the individual's face, creating a vector approximating gaze. Using Euclidean distance, the proximate of the vector to key points was used to determine a likely target for the individual's gaze, such as the other person's face or hands. If a specific target was not identified, the gaze would be approximated as looking toward or away from the opposing individual. The image on the right of Figure 3.4 shows the still-frame from the left with lines drawn using the OpenPose points. The two individuals are connected by their left

Figure 3.2: Screenshot from a PRT training video. The recipient (right) is engaged with the puppet and would be less receptive to instructions from the interventionist (left), indicating an inattentive state. Image from (Virgir05, 2015)

and right hands. An additional vector is depicted extending from the individual's faces to estimate their gaze. The expected gaze target is displayed above each individual. This figure also shows that body points can be overlooked by OpenPose due to occlusion. The individual on the left is missing the points on her left arm and lower body. Her face was also not recognized in enough detail to adequately plot, resulting in a relative scatter of individual points, include two points between the two figures. The right arm of the individual on the right was not detected; however, OpenPose was able to correctly discern the individual's face, as shown by more accurate plotting of points around the head.

Missing body part locations were estimated by looking forward a set number of frames for a value, then back propagating an average value. If a value is not found in the set number of frames, the last known value is used for each remaining frame in the segment. If facial features were not detected, a gaze approximation is calculated using the

Figure 3.3: Screenshot from a PRT training video. The interventionist (left) and recipient (right) are engaged in a joint activity, playing with puppets. This is a demonstration of a shared attentive state. Image from (Virgir05, 2015)

eye, nose, and neck values from the body point set. If those points were also not detected, the gaze target value was set to 'unknown'.

### *3.2.2 Evaluation*

Two evaluations were undertaken to assess the PRT video data and evaluate an approach for detecting attention. The first experiment examined the PRT video data and the use of OpenPose for extracting data from the individuals in the video. The second experiment involved building and training SVM models using the OpenPose data.

Analysis of the videos and the application of OpenPose was conducted prior to exploring the problem of identifying child attention in PRT videos. The basic video attributes are presented in Table 3.2, showing the number of 30 frame segments given each label. The figures regarding the labels indicates that from a human perspective, the

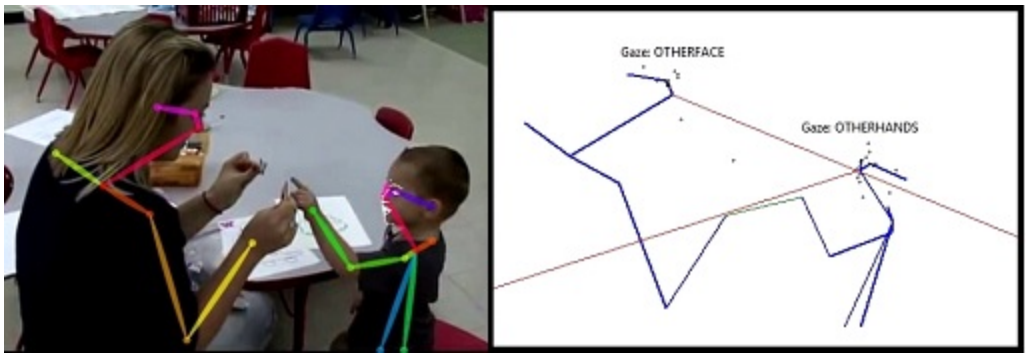


Figure 3.4: The frame on the left is a screenshot from Considine (2011) with an overlay of the face and body points detected by OpenPose. The OpenPose points along with the gaze estimation and hand coordination vectors are shown on the left.

majority of the video segments can be classified. This means that the caregiver and child are visible in the frame, and their interaction is discernible, otherwise the segment would be labeled as 'ignored'. The majority of segments were labeled, indicating that the child and caregiver were generally stationary and easily filmed during their interaction. The ignored segments were not used in the remaining analysis or to train the machine learning models.

Table 3.3 represents the attributes each of the video probes and statistics from using OpenPose. The data shows the ability of OpenPose to extract body and facial features from the videos. These statistics are based both on the obscurity of the individuals in the videos as well as the performance of OpenPose. As expected, OpenPose data displays marginal results. On average only 66% of the body points are identified with an average confidence of 56%. Much of this could be attributed to only the top portion of individuals being in the shot, along with the tendency of caretakers to be on the margins of the video frame. More concerning is the lack of confidence in facial features at an average of 23% recognition confidence. This means that much of the gaze estimation will be undertaken using the less precise locations of the eyes and ears presented in the body point results from OpenPose.

Table 3.2
The attention class label counts for each video probe.

| Video Probe | Attentive | Shared | Inattentive | Ignored |
|---|---|---|---|---|
| Dyad 1 Base | 182 | 43 | 371 | 5 |
| Dyad 1 Post | 170 | 23 | 266 | 156 |
| Dyad 2 Base | 178 | 4 | 254 | 170 |
| Dyad 2 Post | 11 | 585 | 14 | 0 |
| Dyad 3 Base | 146 | 258 | 190 | 10 |
| Dyad 3 Post | 203 | 101 | 133 | 167 |
| Dyad 4 Base | 80 | 0 | 278 | 260 |
| Dyad 4 Post | 261 | 22 | 285 | 33 |
| Dyad 5 Base | 35 | 144 | 415 | 17 |
| Dyad 5 Post | 144 | 66 | 372 | 29 |
| Dyad 6 Base | 95 | 180 | 215 | 125 |
| Dyad 6 Post | 135 | 26 | 317 | 127 |
| Dyad 7 Base | 94 | 110 | 167 | 236 |
| Dyad 7 Post | 119 | 246 | 221 | 24 |

Table 3.3
The proportions and confidence levels for body and facial point detection from OpenPose for each video probe.

| Video Probe | Body Det. | Body Conf. | Face Det. | Face Conf. |
|---|---|---|---|---|
| Dyad 1 Base | 0.72 | 0.58 | 0.76 | 0.13 |
| Dyad 1 Post | 0.62 | 0.55 | 0.6 | 0.08 |
| Dyad 2 Base | 0.62 | 0.56 | 0.87 | 0.36 |
| Dyad 2 Post | 0.69 | 0.59 | 0.8 | 0.35 |
| Dyad 3 Base | 0.63 | 0.5 | 0.85 | 0.26 |
| Dyad 3 Post | 0.53 | 0.56 | 0.79 | 0.33 |
| Dyad 4 Base | 0.74 | 0.57 | 0.79 | 0.23 |
| Dyad 4 Post | 0.74 | 0.56 | 0.83 | 0.23 |
| Dyad 5 Base | 0.72 | 0.57 | 0.95 | 0.34 |
| Dyad 5 Post | 0.72 | 0.53 | 0.78 | 0.17 |
| Dyad 6 Base | 0.55 | 0.54 | 0.63 | 0.1 |
| Dyad 6 Post | 0.59 | 0.52 | 0.63 | 0.13 |
| Dyad 7 Base | 0.62 | 0.55 | 0.82 | 0.24 |
| Dyad 7 Post | 0.68 | 0.59 | 0.91 | 0.32 |

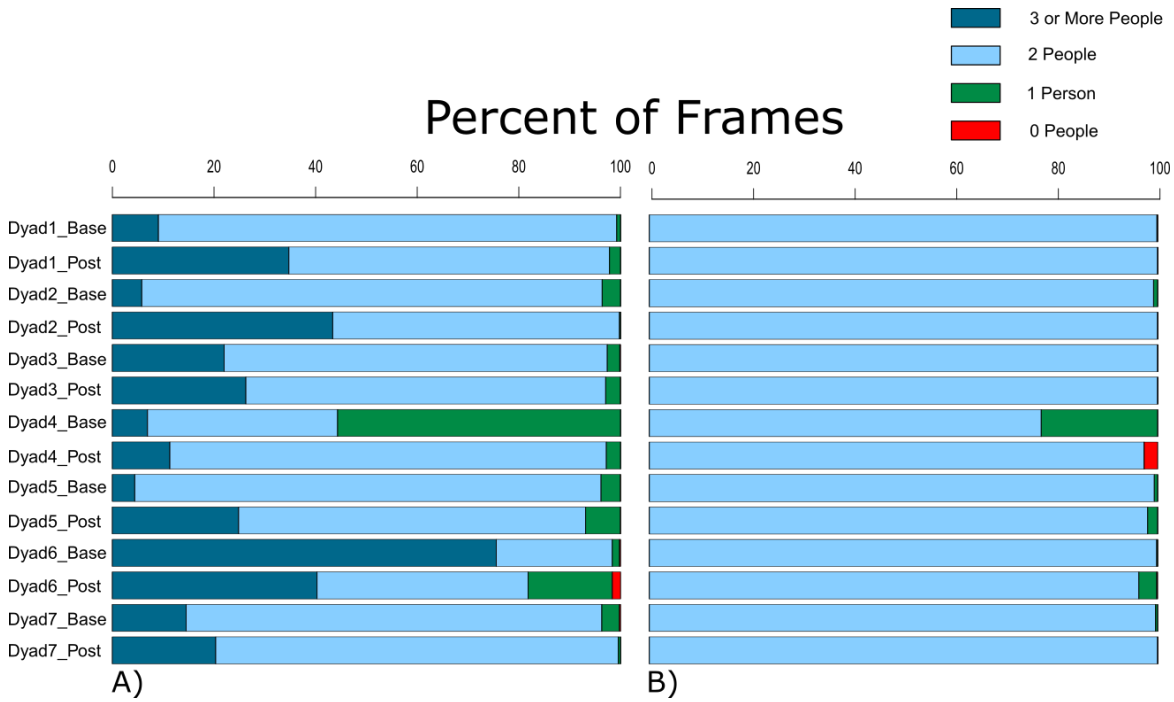


Figure 3.5: A) The bar graph shows the number of people detected by OpenPose in each video. The bars illustrate the percent of frames by the number of people detected. B) Shows the percentages after processing.

Figure 3.5 shows the percentage of frames in each probe by the number of individual people OpenPose identified. The data in the left bar graph illustrates that OpenPose overwhelming recognized only one individual in the frame despite human vetting removing segments where only the child or the caregiver were present. This is likely due to partial occlusion of one of the individuals. In addition to failing to find two people in the frames, there are a significant number of frames where three or more individuals were recognized. This could be due to additional individuals in the background, or to objects being incorrectly recognized as human features.

These statistics were improved by implementing the post-processing procedure for estimating missing data detailed in the methodology section of this paper. The right bar graph of Figure 3.5 shows that this process was able to reconstruct the data to favor having two individuals in the frame. This means that periods of occlusion resulting in the failure

to detect an individual were relatively sparse and the data was able to be approximated within set parameters. This process also caused the completeness of the body points to rise on average across the videos to 74%.

Due to the variety of activities the caregiver and child could participate in, identifying action poses was not feasible and the feature set needed to be generalized. Based on observation and current literature (Koegel, 1988; Suhrheinrich et al., 2011) it was determined that the coordination of the individual's hands and gaze would be the most diagnostic features for detecting attention. Rate of movement was also a consideration, as the child would need to be relatively still in order to be attentive. To emphasize this, the amount of change in the position of body landmarks was included in samples instead of the spatial value. Prior to calculating features, each of the spatial points was rescaled by the frame dimensions of the video and normalized to the neck point of the individual. The position of the child in the frame relative to the caregiver was known prior to processing. This information was used to organize the data so that the caregiver always represents the first person in the feature set.

The feature set used for training the SVM models consisted of a movement score for each individual, the euclidean distance between the hands of the individuals, a gaze target, and a flag indicating reaching behavior. The 30-frame video segments that were labeled were subdivided into six samples. The five frames in each sample where then condensed to provide a summary of the activity depicted over the frames. Each of the six samples were given the same label as the 30 frame segment they were extracted from.

The movement scores were calculated by taking the average magnitude between points in adjoining frames. Unrecognized points were not included in the average. The averages from the frame differences were summed to provide a score for the sample. It is expected that movement scores will be higher when the child is inattentive.

Distances were calculated between the individual's own hands and the hands of the other individual in the frame, resulting in six features. It is hypothesized that when the child's hands are close to one another and far from the caregiver's hands that the child is engaged in an individual activity and is inattentive. If the caregiver's hands and the child's hands are in close proximity, they are likely engaged in a shared activity.

Gaze target estimates are presented as a set of binary features for each frame. This accounts for 10 of the features in the feature set, with five possible targets for each individual. Two target values were based on an individual's gaze intersecting with the other person. One value signified if the individual was looking at the other person's hands or face. A separate value was determined if the individual was looking at the other person's body, but not discernibly at the hands or face. Looking in the direction of the person but not directly at them is the 3rd category, while the 4th category is looking in the opposite direction. The final category is 'unknown' and indicates that the individual's gaze could not be calculated due to missing data points. For the amalgamated feature, the gaze binary values were averaged over the frames in the sample.

The final feature is a flag indicating if the child is reaching. This was determined using the angles between the child's shoulder, elbow, and wrist. In order for the flag to be true in the sample, each of the frames must indicate the child was reaching.

A nu-SVM (Pedregosa et al., 2011) was used to evaluate classification tasks using the dataset. The base and post videos for each dyad used a validation set, while the remaining data was used for training, resulting in seven individual model tests.

Two additional feature sets were used as a baseline to compare the results. The Red Green Blue (RGB) pixel values from individual video frames were used to finetune an AlexNet CNN (Kratzert, 2017) with pretrained weights. Six frames (with every fifth frame starting with the first) from each label segment were selected and ascribed the same

label as the segment they were taken from. The frames were organized in the same training and validation sets are previously mentioned.

The second baseline feature set consisted of the raw spatial values extracted from OpenPose. The data was processed to approximate missing body points and remove additional individuals as mentioned above. Like with the RGB data, this dataset was created using every fifth frame from each labeled 30 frame segment and organized into the same training and validation sets.

## 3.3 Results and Discussion

The results of the model evaluations reveal the complexity of the problem space. Overall accuracy for each approach is low and varies substantially between validation sets as depicted in Table 3.4. In the table, the accuracy metric shows the proportion of correct to incorrect predictions on the validation dataset. The segment accuracy is classification of the 30 frame segments that were originally labeled. This score is determined by taking the majority of the predicted class for the six samples that were classified for each segment.

Table 3.4
Proportion of correct label predictions for RGB, Spatial, and Expanded data for each validation set.

| Evaluation | Dyad 1 | Dyad 2 | Dyad 3 | Dyad 4 | Dyad 5 | Dyad 6 | Dyad 7 |
|---|---|---|---|---|---|---|---|
| AlexNet RGB Accuracy | 0.46 | 0.39 | 0.37 | 0.55 | 0.55 | 0.40 | 0.33 |
| AlexNet RGB Segment Acc. | 0.49 | 0.37 | 0.37 | 0.59 | 0.59 | 0.52 | 0.57 |
| SVM Spatial Accuracy | 0.32 | 0.31 | 0.32 | 0.43 | 0.4 | 0.34 | 0.34 |
| SVM Spatial Segment Acc. | 0.41 | 0.53 | 0.60 | 0.51 | 0.47 | 0.50 | 0.62 |
| SVM Expanded Accuracy | 0.43 | 0.37 | 0.42 | 0.51 | 0.50 | 0.49 | 0.41 |
| SVM Expanded Segment Acc. | 0.52 | 0.51 | 0.54 | 0.6 | 0.63 | 0.57 | 0.56 |

The low average accuracies show that there is a significant variation in the data, making generalization of patterns for classification needed for successful predictions difficult. The dataset is strongly imbalanced, favoring inattentive behavior. This was

addressed by undersampling the larger classes to be equal size with the smallest. This resulted in a reduction of the samples that could be used to train the models. Additionally, samples only needed to display attention for part of the segment to be classified as attentive. This could lead to subsamples of the attentive segments being similar in composition to inattentive or shared attention samples.

Related to the data imbalance in the dataset, the video probes do not have equivalent class compositions, making training a model that performs adequately on each set difficult. This is particularly apparent in Dyad 2 and Dyad 7, which have large quantities of shared attention samples. The shared activity largely depicted in the Dyad 2 post video is the child sitting in the caretaker's lap watching a movie on a cellphone. This is distinctly different from other shared activities that often had the child and caregiver facing each other, with the motivator, a toy or game, in between them. Similarly, in the Dyad 7 post video the child and caregiver are sitting across the table from one another playing a game. The distance in between them is much greater and there are longer periods of inaction than in other examples of shared activities.

The worst performance was exhibited by the spatial feature set at 35% average accuracy, which is not significantly different from what would be expected with random label assignments. As this data only contained the coordinate location of the individuals, there was little for the algorithm to generalize to form an effective classifier. The inclusion of this feature set was to provide a baseline to compare the results from the CNN model and processed feature set.

The AlexNet CNN performed better than the spatial feature set at an average of 43% accuracy across the validation sets. This illustrates that the pretrained network was able to extract features from the still frames to improve classification above random; however, the data still exhibited significant variation to prevent strong predictions. The image background alone does not likely account for the variation. Each video was filmed in

rooms, often the same room, at the same ASD resource facility. As such, the children in the videos often had access to the same toys. Since the toys would be associated with different states of attention, this could have aided the classification.

The intention of the processed dataset was to generalize the features of each state of attention to provide a classification regardless of the activity. At an average of 46% accuracy, this was not sufficiently achieved. As mentioned above, the variability among activities caused some scenarios to be in the validation set that did not have an adequate equivalent in the training set. More data may be needed to ensure that a wider range of activities is encompassed in both sets.

The precision of OpenPose could also be potentially problematic. Although it has been demonstrated that OpenPose adequately detected the individuals in the frames, the data presented in Table 3.3 illustrates that prediction confidence levels are low, particularly regarding facial features. Since the processed feature set is reliant on extracting information from these points, it is important that recognition be uniform for each sample. The detected position of body parts could vary between videos based on the quality of the video, the proximity of the individual to the camera, whether the person has his or her back to the camera, or other external factors such as clothing. As each of the videos are taken by a handheld device, these issues will be compounded by the movement of the camera. Each of the 30 frame increments in the video were processed through OpenPose independently. It is likely that processing the video as a whole would improve the precision of detection through OpenPose's tracking algorithms.

The two-dimensionality of the data was problematic for detecting the target of the individual's gaze. One problematic scenario that arose in several instances occurred when the child was playing with a toy, a ramp set for cars, while the parent stood behind the child and watched. The child gaze is on the toy; however, because of the parent in the background the child's gaze would be incorrectly attributed to looking at the parent.

Similar issues regarding incorrect gaze targets can also be attributed to a lack of the ability to infer eye direction. For instance, a child facing the camera, but looking down at the cellphone in his caregiver's hand, is in a shared state of attention, whereas a child looking directly at the camera, or interacting with the person filming the scene, is not attending the parent. The difference in head position is not discernible from the OpenPose data causing misidentification of gaze. This inadequacy may be addressed by incorporating RBG data of the individuals' heads as a substitute for attempting to determine a gaze target.

The ultimate goal of this research is to maximize the feedback that can be provided to the PRT practitioner while reducing the amount of time necessary for expert review of the video probes. Determining the child's attentive state is an important part of PRT. Even with the low classification accuracy, this system could provide benefit to a clinician reviewing the videos. The segment accuracy score presented in Table 3.4 represents when the system correctly detected the child's attention to the caregiver in regards to when the caregiver can provide a prompt or instruction. As such, shared and attentive states have been combined. Table 3.5 contains a description of two minutes from the Dyad 5 post video probe along with the attention classifications. Each second of the video is classified. The predictions are correlated with the activities in the video, and show that the child was largely engaged in solo play with intermittent periods of attention toward the caregiver. By providing this information to the clinician he or she could gauge the relative attention of the child throughout the video along with finding moments where the caregiver should be providing instruction. This also provides a metric that can be viewed across multiple videos to gain an understanding of whether the caregiver is improving at seizing the child's attention.

This example also illustrates that in the scheme of attention, accuracy to a one second precision is unnecessary. This research presented above examined detecting attention in independent one second segments. A more robust approach will likely need to account for

greater lengths of time and encompass a broader understanding of the temporal features for detecting attention. This is particularly important for shared attention states as they are generally sustained for a long period of time compared to attentive states.

Table 3.5
Illustration of how attention classification aligns with behavior in the video probes. The first two minutes of Dyad 5 post video probe has been broken into 30 second increments with each second classified as attentive (attn) or inattentive (inattn).

| Time (secs) | Description | Attn | Inattn | Acc. (%) |
|---|---|---|---|---|
| 1 – 30 | The child looks about the room while the caregiver offers a choice of toys. The caregiver offers a different choice of toys and the child responds | 13 | 17 | 63 |
| 31 - 60 | The child plays with the toy. The caregiver offers a choice of accessories. The child chooses and continues solo play. | 8 | 22 | 70 |
| 61 – 90 | The caregiver offers a new choice of accessories. The child chooses, but then asks for a different piece. After receiving the piece, he continues playing. | 8 | 22 | 83 |
| 91 - 120 | The child continues solo play. Then watches the caregiver rummage through accessories. The child is offered a choice but does not respond correctly and the caregiver holds the toy until a new response is given. | 12 | 18 | 93 |

## 3.4 Conclusion

Video probes are an integral part of evaluating people learning PRT. The current manual process is limited by the costs of having behavior analysts extract relevant data. These videos provide an opportunity for both for improving the training process for learning

PRT as well as expanding the field of dyadic human activity detection and classification. PRT video probes provide a complex problem for computer vision due to the unpredictability of the camera stability, the mobility and occlusion of the individuals in the video, and the range of activities that could be performed. The research examined three data representations using two different machine learning algorithms to detect dyadic attention in untrimmed videos to serve as a baseline for future research. Greater exploration into extracting important diagnostic and temporal features is needed to improve classification predictions.

# Chapter 4

# VOICE ACTIVITY DETECTION AND PARENT-CHILD SPEAKER SEPARATION

Detecting vocal activity in PRT video probes is difficult. Often, the audio tracks contain background noise from the environment along with sounds from the activity the child and parent are participating in. This could include sounds from play activities, toys that emit songs, chimes, or speech recordings, and dialog from electronic media. These noises can obscure the parent or child vocalizations or create opportunities for misidentifying a speech event. The recording quality of the child and parent can also be problematic, as the videos are often recorded using handheld phones or cameras with built-in microphones. The quality is therefore dependent on proximity to the camera's microphone. This is particularly limiting for children with low energy vocalizations.

An additional challenge, and what distinguishes this research from other works on voice activity detection (VAD), is that the parent and child exhibit atypical speech patterns. To engage the child, the parent often utilizes child-directed speech patterns, or baby-talk, which involves drawing out syllables and using a higher pitched voice in a way that is not common in adult speech. Child speech is already a difficult problem for automatic speech detection (Lee et al., 1999), as children speak more slowly than adults and make more phonetic or grammatical errors. This could be more prevalent in children with ASD who have limited communication skills. Additionally, in PRT, a valid vocalization from a child is determined by his or her communication ability. This means that a child who is non-verbal or whose speech is limited to single words may only be able to respond with a phoneme in response to a learning objective. Because of this, it is important to detect all the child's vocalizations, not just articulated speech. The research presented below evaluates methods for detecting parent and child vocalizations in PRT video probes.

Processing audio for speech-related tasks is a well-studied area of computer science. Approaches vary extensively throughout the literature, differing in the ultimate tasks being undertaken, the models that are employed, and the feature sets that are used for detecting and classifying speech signals. For the PRT project, it is relevant to discuss feature extraction, VAD, speaker separation, automated speech recognition (ASR) systems, and nonspeech vocalizations. Additionally, the application of these approaches to children, particularly children with ASD, is important. After a discussion of the relevant literature, the audio data corpus created from the videos are analyzed and research into applying audio processing and classification methodology for PRT video probes is discussed.

To evaluate the audio signals from the PRT video probes, several detection methods were examined, including filter-based implementation, clustering algorithms, and machine learning approaches. These results are compared to the open source VAD system, WebRTC VAD[1].

## 4.1 Research in Audio Processing

Several facets of audio processing are applicable to this project. The most important concepts for the current implementation are VAD and speaker separation. An examination of acoustic features is also important for understanding what aspects of the audio signal are going to be diagnostic and robust when addressing the problems inherent to PRT video probes. Additionally, detecting non-speech vocalizations and unvoiced speech has a correlation to detecting child utterances. Automated speech recognition (ASR) and how noise is handled in ASR models is relevant; however, as language recognition is not currently addressed in the project, these subjects only warrant an introduction in this chapter.

[1] https://webrtc.org/

### *4.1.1 Acoustic Feature Extraction*

Audio signal processing is based on three primary types of data features: spectral, cepstral, or prosodic. Spectral features represent the times series of audio wave frequencies in accordance with a central tendency, spectral centroid, and the periodicity of change, spectral flux. Cepstral features are created by transforming spectral features. Common transformations used in speech analysis include Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC), and Perceptual Linear Prediction (PLP) (Dave, 2013). A formant is a spectral or cepstral pattern associated with a sound such as a musical note or a phoneme. Prosodic features are not associated with formants and include rhythm, intonation, and stress. Prosodic features are often used in transcription tasks to signify punctuation or meaningful boundaries in natural speech (Moore et al., 2016) and in emotion detection. Common feature extraction tools in research publications include PRAAT (Boersma and Weenink, 2018), GeMAPS, OpenEar (Eyben et al., 2009), OpenSMILE (audeering, 2018), and WaveSurfer (2018).

Developed to emulate anatomic hearing, PLP features focus on the spectral configuration of the audio signal (Hermansky, 1990). This analysis is based on selecting a temporal window size for segmenting the signal, then transforming the raw data using a fast Fourier transformation (FFT). This translates the wave information into a frequency measurement.

Although seen as effective features for speech recognition (Hönig et al., 2005), PLP could be susceptible to interference based on recording quality or background noise (Hermansky, 1990). In relation to this, PLP features were seen to be important for detecting speech intelligibility based on regression models (Salehi and Parsa, 2016).

LPC is a methodology for compressing audio by approximating the signal using a linear prediction model (Dave, 2013). The signal representation can be effectively used in

speech activity detection and recognition tasks. This reduces the feature space of the signal, aiding in more robust classifications (O'Shaughnessy, 1988).

Like PLP, MFCC was designed to replicate features of anatomical hearing. This feature extraction approach focuses on using the Mel scale, which is based on the frequency range of human hearing (Dave, 2013). Creating MFCC features is based on selecting an arbitrary segment size and performing FFT to translate the time-series dependent wave into a frequency-based feature. As MFCC, like PLP, is based on the spectral signal, speech recognition models using the features will be sensitive to noise (Shrawankar and Thakare, 2013).

Zero crossing rate and energy are often used in discerning speech from non-speech in audio signals (Bachu et al., 2008, 2010; Shete et al., 2014). ZCR is a useful feature for speech identification tasks because it is largely independent of speaker energy (Ito and Donaldson, 1971). The ZCR of the signal represents how frequently the signal values change from positive to negative, or vice-versa. The pattern in the crossing frequency can be indicative of specific phonemes in natural speech patterns. Calculating ZCR is a straightforward count of transitions over the zero-center point in a specific interval of the signal pertaining to an amount of time. The expectation is that the ZCR will be low of audio segments containing speech compared to non-speech segments (Shete et al., 2014). In addition to ZCR, cepstral peak can be a diagnostic feature for discerning unvoiced speech (speech sounds that do not use vocal cords) (Graf et al., 2015).

The energy of the signal is related to the amplitude and modulation. Energy has been seen as an important acoustic feature for discriminating voice signals from other audio. This is based on peak amplitude patterns in speech signals differing from unvoiced signals. The expectation is that energy amplitude in records will be higher for audio segments containing speech compared to segments with no speech (Shete et al., 2014).

### *4.1.2 Voice Activity Detection*

Voice activity detection (VAD) encompasses the preprocessing techniques for discriminating speech signals from other noises in an audio file. Generally, approaches to classifying speech versus non-speech signals involves using discriminatory feature sets, statistical approaches, or machine learning techniques (Zhang and Wang, 2016). A common feature-based technique is the use of frequency ranges as a filter for selecting speech signals (Aneeja and Yegnanarayana, 2015; McLoughlin, 2014). Statistical approaches focus on modeling the noise spectra using a defined distribution in order to extract impertinent signals.

Both unsupervised and supervised machine learning methods have been explored for VAD. In unsupervised methods, k-means (Górriz et al., 2006) and Gaussian mixture models (GMM) (Sadjadi and Hansen, 2013) have been explored. Unsupervised methods benefit from the ability to use large amounts of data; however, the algorithms falter in difficult separation tasks, such as when a noise signal has a steady repetition (Zhang and Wang, 2016).

Support vector machines (SVM) have been a commonly utilized algorithm for VAD (Enqing et al., 2002; Jo et al., 2009; Shin et al., 2010). These approaches focus on utilizing the SVM for a binary classification problem, requiring a labeled corpora of noise and speech data. The requirement for label data is the primary drawback for these approaches, particularly due to the variety in noise and speech signals. This means that the model may not be able to generalize to compensate for different types of noise.

Deep learning approaches for VAD seek to address generalization by utilizing the network layers to capture more information about the data's feature set. The use of a feedforward recurrent neural network (RNN) model for VAD was explored by Hughes and Mierle (2013). A single hidden layer neural network was implemented by Drugman et al. (2016) and was applied to test VAD application in real world environments. Also

exploring application to real world scenarios, Kim and Hahn (2018) utilized multiple layers of encoder and decoder to networks to create a classification model.

### *4.1.3 Speaker Separation*

The "cocktail party problem" (Cherry, 1953) examines the task of separating voices from a single audio track. This is based on the human ability to focus on individual voices in a cacophony. Often, research addressing the cocktail party problem involves identifying individuals speaking simultaneously. Approaches for speaker separation have focused on the use of similarity metrics.

Morgan et al. (1997) utilized an approach based on identifying signatures near the start of the signal, then inferring similarity using likelihood metrics in subsequent segments. They examined the frequency strength of the signal to determine a strong and weak signal. Based on these two channels, subsequent signal segments are analyzed to determine which classification they belong to. Differentiating modulation was also examined for speaker separation by Schimmel et al. (2007). Similarly, Yu et al. (2017) researched identifying optimal output assignments using a CNN. Their network models would perform the separation based on training the network to minimize the error when comparing sample magnitudes.

Clustering and masking algorithms examine signal features in order to assign classification based on similarities in comparison. Masking approaches focus on discovering identifying features of the dominant speaker, then setting other components of the signal to an insignificant value (Reddy and Raj, 2007). These features are identified using grouping methods. Similarly, clustering approaches (Chen et al., 2017b; Isik et al., 2016) examine temporal segments of the signal and utilize a comparison algorithm to separate the segments into distinct classes. Clustering approaches benefit from being unsupervised, and do not require additional data for training. However, some approaches

use training data for implementing neural network models that can be used to reduce the feature space as in the case of deep clustering (Isik et al., 2016).

### *4.1.4 Automated 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 soundwave as a time-series signal, either transforming the signal to isolate discriminative features for classification (Dave, 2013) 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 (Abdel-Hamid et al., 2014; Abushariah et al., 2010; Graves et al., 2013; Hinton et al., 2012; Jaitly et al., 2012; Rabiner, 1989; Sak et al., 2014; Wilpon et al., 1990). These implementations utilize a hybrid approach implementing a classification model for identifying formant or signal segments, then using an additional model such as a hidden Markov model (HMM)-Gaussian mixture model (GMM) combination to infer temporal relationships.

End-to-end speech recognition refers to architectures that do not use this hybrid model, and simultaneously perform feature and temporal classification tasks with a single model, often a DNN-based model, a CNN (Sainath et al., 2015; Zhang et al., 2017a), or RNN (Chan et al., 2016; Chen et al., 2016b; Miao et al., 2015; Sainath et al., 2015). Similar end-to-end implementations have used a deep belief network (DBN) model based on restricted Boltzmann machines (RBM) (Dahl et al., 2012; Sarikaya et al., 2014).

ASR implementations using raw data, colloquially called end-to-end speech recognition, use the raw wave data from the audio signal instead of transforming the signal into frequency space, as in PLP or MFCC space, or extracting other features. These approaches have focused on identifying phonemes using a DNN or CNN implementation

(Golik et al., 2015; Hoshen et al., 2015; Palaz et al., 2013, 2015a,b; Passricha and Aggarwal, 2018; Tuske et al., 2014).

### *4.1.5 Addressing Noise in Speech Recognition*

Depending on the environment, different degrees of noise can be expected for ASR applications, making noise-robust approaches an important part of ASR research. There are several options for creating a more robust ASR system that are involved at both the feature and the model level. Feature-based methods for addressing noise focus on utilizing extracted features that are inherently robust or implementing preprocessing procedures to add robusticity. PLP features typically are naturally robust to noise. Preprocessing procedures predominantly use normalization (Li et al., 2014). Model-based approaches incorporate adaptation and compensation for specific noises.

Having prior knowledge of the audio track recording environments allows for compensation of specific noise patterns. The expectation is that certain noises will be prevalent in the environment and can be compensated for during feature selection and model training(Kim and Hahn, 2018).

Training models on deliberately distorted speech can help generalize classification tasks. During training, distortion is added to clean samples in order for the model to gain flexibility. Typically, this is undertaken using statistical or sample-based methods for determining the distortion effect. A probability-based example was presented by Bu et al. (2018). Their approach involved augmenting the spectral features using their temporal context.

Uncertainty-based approaches examine the likelihood the model is correct, and its confidence in its predictions. This can occur at both the model or feature level. This can be approached by implementing error correction in a long-term training context using DNN implementations (Shivakumar et al., 2018).

Joint training methods incorporate data with noise into training the classification models. This approach was used by Narayanan and Wang (2014). They trained a DNN model on speaker separation tasks as well as speech recognition. Their intuition was that this would make the model more powerful at discerning feature discrepancies and increase generalization.

#### *4.1.6 Child Speech Recognition*

Performing speech recognition on children presents additional challenges to automated speech processing. At an auditory level, children's voices tend to be higher frequency and display more rational and spectral variability (Lee et al., 1999). Regarding language modeling, children are more prone to mispronouncing words than adults, have a restricted vocabulary, and tend to speak at a lower rate (Potamianos and Narayanan, 1998). These challenges are more apparent the younger the child is. Research into child speech classification has been undertaken using SVM models (Boril et al., 2014), DNN models (Dubagunta et al., 2019; Liao et al., 2015; Ward et al., 2016), and hybrid DNN – hidden Markov model (HMM) classifiers (Smith et al., 2017). Discerning adult from child speech was explored in (Aggarwal and Singh, 2015). Adding adult speech samples when training child speech recognition models has been shown to improve classification accuracy (Smith et al., 2017; Ward et al., 2016).

In dyadic speech classification, domain adaptation and the utilization of contextual information implemented were used to increase recognition accuracy by Kumar et al. (2017). Their system examined speech from child-adult interactions in child mistreatment interviews using separate networks for the adult and child speech recognition. Domain adaptation on the children's speech network consisted of incorporating transcripts in training to aid in structuring the data. Additionally, the researchers sought to use the recognized adult speech as context to infer more accurate transcription of the speech from

the child in the interaction. Using this approach, they showed that substantial improvements in word recognition accuracies were made in comparison to a baseline measurement.

### *4.1.7 Phoneme and Vocal Event Recognition*

In addition to speech detection, non-speech vocalization is also important to detect. Detecting laughter and filler utterances, such as 'um', 'er,' or 'eh' vocalizations, was the focus of work conducted by Gosztolya (2016). To accomplish this, the researchers investigated DNN and AdaBoost models utilizing smoothing algorithms for aggregating probabilities along the audio time series. Their classifier was based on three classes: laughter, filler, and other, with the other category encompassing silence and fully articulated speech. They found that applying smoothing algorithms improved the performance for discerning each class.

Examining detection of native language and deception, Gosztolya et al. (2016a) compared the results between SVM, DNN, and AdaBoost classifiers. The classification tasks were conducted on phoneme features extracted from the speech data. Their results showed that using a combined architecture of a DNN and AdaBoost produced the best performance in terms of accuracy and recall on native language detection. On detecting deception, their approaches did not surpass the baseline SVM implementation (Schuller et al., 2016) in accuracy; however, the DNN model achieved better recall.

SVM models were used in a study examining automatically created feature sets for detecting if a speaker in audio data has a mild cognitive impairment (Gosztolya et al., 2016b). Their work provides insight into the most diagnostic features for detecting cognitive impairment. Their results show that demographic information was not a significant factor in detection. Speech and articulation rates, and utterance length, were also not significantly present in the feature sets with the highest performance metrics.

Non-speech related phonemes were a major inclusion in the successful feature sets. These are important because they can represent hesitation in speech. The authors state that despite this importance, these features are often overlooked in cognitive impairment classification.

Research into child pronunciation was conducted by Dudy et al. (2017). In their publication, the authors' goal is to create a system that can be utilized to detect mispronunciations for children that have speech disorders in order to create a system that can be used as an aid for improvement. Two models were used to classify pronunciation - one based on an SVM architecture and the other utilizing a GMM-HMM. They concluded that both methods had similar performance, with the GMM-HMM implementation having a slightly better accuracy.

Sentence detection is an important part of automated transcription and conversation analysis tasks. Delimiting speech is a difficult problem in spoken language as people often restart sentences, pause at random, leave sentences unfinished, or get interrupted (Moore et al., 2016). In order to break points in speech, Moore et al. (2016) trained an SVM on prosodic and lexical features. The output of the SVM was used to create a probability that the classified speech unit was a breakpoint. They concluded that the model predominantly utilized the prosodic features. Removing the lexical data from the feature set caused little change in the classification accuracy of the model.

#### *4.1.8 Application to Autism Research*

Much of the research regarding the implementation of ASR systems for individuals with autism has focused on diagnosis and emotion detection. Exploration of the application of ASRs for emotion detection in children with autism was undertaken by Marchi et al. (2015). The dataset consisted of both children with autism and children without who were acting out emotions based on story prompts. Classification of emotion class was

undertaken using an SVM. Their findings indicate that larger feature sets equated to better performance. They found that system had a higher detection recall rate for the children without an ASD diagnosis.

Researching Autism detection, Xu et al. (2009) 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 processing time for evaluating language skills for people with autism. After recording the audio, the system sought to classify the vocal data into classes, including the target child, adults, other children, and voices from electronic media using a GMM-HMM model with a high dimensional set of features (Xu et al., 2008). They concluded that their work illustrates a high degree of difference in speech between children with autism and children without a diagnosis of ASD that can be suitably differentiated using machine learning techniques.

The LENA recording system was also used by Pawar et al. (2017) to analyze vocalizations of children with autism and their interaction with adults. Their approach utilized a SVM classifier to distinguish between adult and child utterances as well as detect laughing. Their results were comparable with Xu et al. (2009, 2008).

This project differs from much of the work on VAD and speaker separation because of its implementation in handling adult and child vocalizations, along with unpredictable noise. Additionally, the project needs to account for children with limited verbal skills that may not be able to formulate complete words and adequately recognize all vocal utterances.

The LENA system provides a similar function to the research presented in this paper. This paper focuses on classifying audio from untrimmed videos of PRT sessions. This is intended to work within the current structure of PRT implementation and research practices. The videos can be unpredictable in the interactions depicted and the quality of the recording. The LENA system benefits from using hardware attached to the child's

clothing. This likely provides higher quality recordings, particularly of the child's vocal utterances; however, it is dependent on a specific device.

## 4.2 Corpus Description

The corpus used for audio research was extracted from the video probes described in Chapter 3. Each of the parent-child dyads consisted of an adult female and a male child. The children ranged in age from 24 to 60 months. The communication skills exhibited varied depending on the child. Table 4.1 provides an observation of the vocal abilities that the child shows in each of the videos. The majority of the child vocalizations expressed do not consist of fully articulated words. The data from five of the seven children contained few single word utterances. The child utterances in these videos primarily consists of sounds unrelated to speech or attempts to pronounce the first phoneme of a prompted response. The child from the Dyad 4 videos spoke in single words, or two word phrases with some additional non-speech vocalizations related to play activities. The child in Dyad 7 spoke in multi-word phrases with few non-speech utterances.

Table 4.1
Speech level exhibited in the video probes by the child in each dyad.

| Child | Exhibited Vocal Skill |
|---|---|
| Dyad 1 | Vocal attempts, single words |
| Dyad 2 | Vocal attempts, no fully articulated speech |
| Dyad 3 | Vocal attempts, single words |
| Dyad 4 | Single words, two-word phrases |
| Dyad 5 | Vocal attempts, single words, two-word phrases |
| Dyad 6 | Vocal attempts, single words |
| Dyad 7 | Multi-word phrases, full sentences |

The parent speech consists of individual words, sentences, and exclamations. Much of the parent's speech follows child-directed speech patterns. This consists of using a higher pitch than in normal conversational speech, along with extending syllables and

exaggerated excitement or surprise. Only the parents vocal utterances were attributed to the adult in the labeled audio segments. Sounds made by the parent that were not verbalization, such as clapping or sneezing, were labeled as noise.

In the video, various play scenarios are participated in, creating different types of noises including shuffling toy pieces and objects banging together. Additionally, the toys themselves often emitted noise, such as a dinosaur roar, music, or audible speech. In one video, Dyad 2 Post, the parent and child are watching a popular children's movie on a mobile phone. Speech from the toys or other media were omitted from the dataset. Sounds from the movie that were not recognizable speech were labeled as noise.

In addition to the parent talking in the video, there are some instances of an additional adult in the room speaking. For this publication, only audio from the parent is used in the dataset. Audio segments were labeled at 250 ms as either parent speech, child vocalization, or non-speech sounds. Segments with an energy level below $1e^{-6}$ were excluded. The number of labeled segments for each video are posted in Table 4.2.

Table 4.2
Number of labeled samples for each of the three classes for each video probe.

| Video | Parent Vocalization | Child Vocalization | Non-speech Audio |
|---|---|---|---|
| Dyad 1 Base | 797 | 591 | 1049 |
| Dyad 1 Post | 156 | 162 | 622 |
| Dyad 2 Base | 763 | 120 | 1533 |
| Dyad 2 Post | 365 | 64 | 700 |
| Dyad 3 Base | 1017 | 124 | 1208 |
| Dyad 3 Post | 477 | 247 | 1645 |
| Dyad 4 Base | 1358 | 375 | 702 |
| Dyad 4 Post | 967 | 429 | 1009 |
| Dyad 5 Base | 705 | 97 | 1538 |
| Dyad 5 Post | 509 | 248 | 1686 |
| Dyad 6 Base | 574 | 108 | 1778 |
| Dyad 6 Post | 295 | 132 | 1996 |
| Dyad 7 Base | 923 | 785 | 708 |
| Dyad 7 Post | 797 | 591 | 1049 |

## 4.3 Experiments and Results

The objective of the experiments is to find an algorithm that is able to detect vocalizations in the video, determine if they are from a child or an adult, and to identify noise segments. To achieve this, methods incorporating WebRTC VAD, pitch-based filtering, clustering algorithms, and machine learning techniques were explored.

The first experiment that was conducted was to determine how well a state-of-the-art VAD system performed on the video probe data. Google's WebRTC VAD [1] is an open-source tool for extracting speech segments from audio files. Each of the video probe files was processed using WebRTC VAD independently. The VAD can be configured to integer-based levels of aggression that influence the threshold for determining noise from valid speech. The two lowest levels of aggression, one and two, were tested. The results are presented in Table 4.3.

Table 4.3
The percent of label segments that were correctly included in an audio segment if a vocalization or excluded if noise after processing each video with WebRTC VAD.

| Aggression Level | Ave. Correct Noise | Ave. Correct Adult Speech | Ave. Correct Child Vocalization |
|---|---|---|---|
| One | 0.27 | 0.98 | 0.97 |
| Two | 0.91 | 0.12 | 0.06 |

The results show that WebRTC cannot accurately filter the video probes. On the lowest setting, the majority of vocal samples were correctly captured by the VAD; however, noise was not sufficiently filtered. On this setting, 73% of the noise samples were included in the processed audio segments. Conversely, on aggression setting two, 91% of the noise was correctly removed, but the majority of speech samples were not captured, particularly for the child utterances. This performance is likely due to several

[1] https://webrtc.org/

factors. The VAD may be designed to filter environmental noises which may be periodic or droning, and is thus looking for anomalous signal magnitudes to detect speech events. The noise in the video probes does not fit this pattern and is usually the result of the child or parent playing with a toy or participating in an activity. Detecting noise may also be based on energy levels. The noises in the video probes are often high energy events whereas the vocalizations, particularly from the child, may be low energy.

The second experiment sought to distinguish between noise, child vocalization, and adult speech using a filter on the estimated signal pitch for each 250 ms segment. The estimated pitch was extracted using PRAAT (Boersma and Weenink, 2018) within a range of 75 to 600 Hz. The average estimated pitch for the segment was calculated and used for classification. The classification model used a rule based on the expected average range for female adults and male children. The range for adults was 165-255 Hz (Titze and Martin, 1998). The child range was 260-440 Hz, based on information from Hunter (2009). The results are illustrated in Figure 4.2. This method had marginal success in determining noise segments, with an average F1 score of 80%. Adult and child segments were less successful, with average F1 scores of 52% and 39% respectively. This shows that much of the noise in the segments falls outside of the pitch range of 165 to 440 Hz. It is also notable that the method had the best success in classifying child vocalization in Dyads 4, 6, and 7. These children exhibited more complete word usage.

Recorded pitch frequencies for children in research studies is varied (Hunter, 2009). In the corpus presented in this study, both the child vocalizations and the adult speech registers at a higher estimated pitch than other publications. Figure 4.1 presents a box plot for the average estimated pitch frequencies for adult, child, and noise segments for each video. The range of all three classes extends from 75 to 600 Hz based on the parameters provided to PRAAT. This indicates that samples in the adult and child classes contain samples outside the expected vocal range. The means of both are higher than reported in

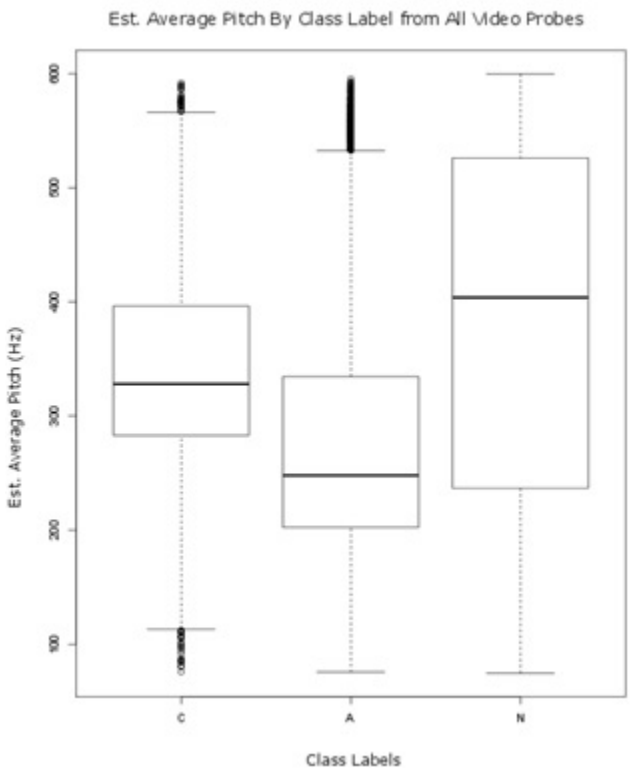


Figure 4.1: Box plot showing the average estimated pitch for segments in the video probes. The plot shows the distribution for class labels C, A, N, or child, adult, and noise respectively.

other publications. For child samples, the mean is 343 Hz and for adult samples it is 279 Hz. The means of each class is distinct; however, the interquartile range shows a large degree of overlap.

The estimated pitched-based classifier described above was rerun using ranges from the dataset distribution. The adult and child ranges were based on the 1st and 3rd quartiles. The region of overlap between the parent and child data was handled by dividing the region and ascribing samples in the higher frequencies to the child. This gave an adult range of 202 - 308 Hz and a child range of 308 - 396 Hz. The results are compared to the previous implementation in Figure 4.2. This method gives a narrower range of values for the adult and child classes and exhibited a lower accuracy than the previous method, based on published frequencies. This discrepancy likely shows that

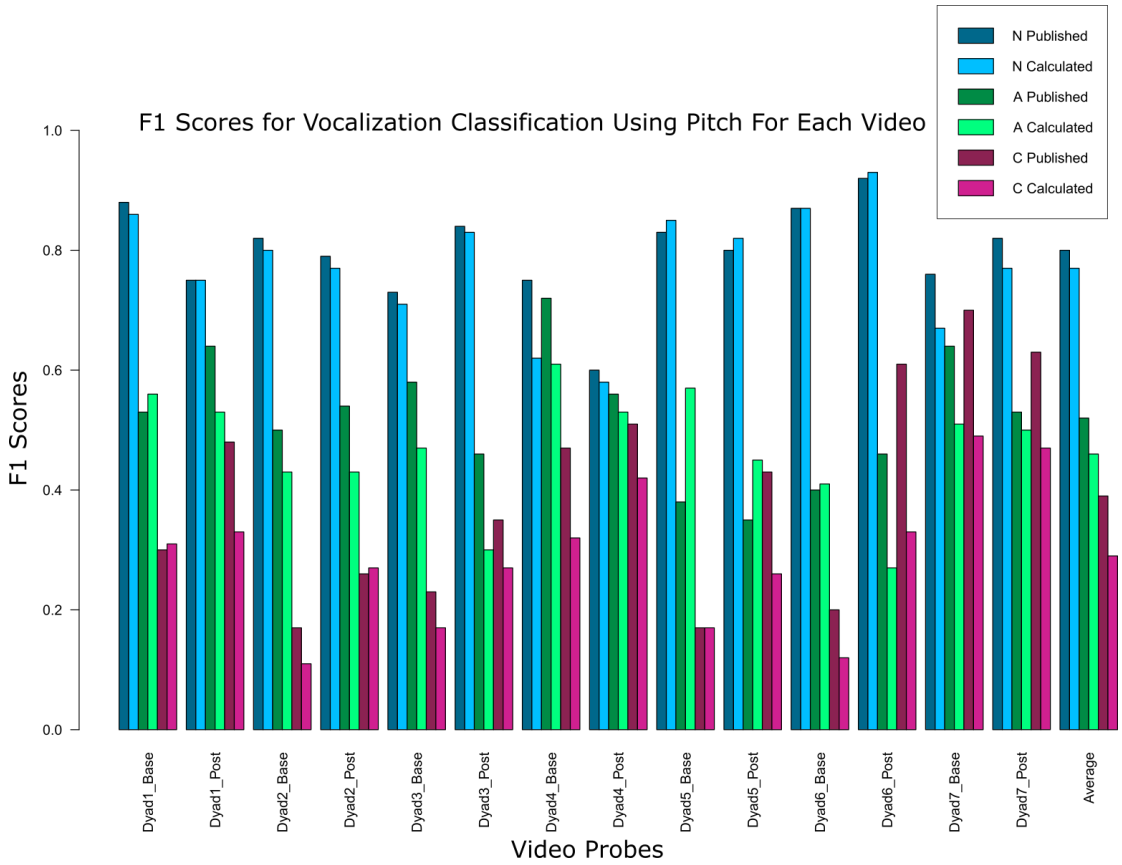


Figure 4.2: F1 scores from using estimated average pitch to classify audio from the video probes. Two value ranges were explored on the dataset with the child range of the first classifier being 260-440 Hz and the second classifier being 308-396 Hz.

outliers in the data are skewing the frequencies. This could be due to variance in the energy of samples causing less accurate estimates of the pitch quality. It could also be the case that exclamations and exaggerated excitement could cause the adult pitch estimations to be higher than spoken language.

The third set of experiments utilized the open-source library PyAudioAnalysis (Giannakopoulos, 2015) for feature extraction and running machine learning algorithms. This experiment compared five classifiers that are available in PyAudioAnalysis: support vector machines (SVM), k-nearest neighbors (KNN), random forests, extra trees, and

gradient boosting. Each of these classifiers is implemented with the sklearn python library (Pedregosa et al., 2011).

For processing, each labeled 250 ms segment was saved to a wav file. The wav files were converted into 68 element vectors consisting of the midterm features extracted by PyAudioAnalysis. The feature vectors consist of values for zero cross rate (ZCR), energy, energy atrophy, spectral spread, spectral flux, spectral runoff, mel-frequency cepstrum coefficients (MFCC), chroma, and chroma standard deviation. The feature set is then standardized prior to training the classifier.

Twelve of the 14 videos were used for training each classifier. The remaining two videos, the base and post video for a single dyad, were used as a validation set. The average results across all validation sets for each model are displayed in Figure 4.3. These results are similar across each of the classifiers, with gradient boosting and SVM providing the best F1 scores for each class. These results also mirror the filter-based results. This shows that the noise segments are easily distinguishable from the other classes, but the human vocalizations are more difficult to classify.

The results from the PyAudioAnalysis algorithms illustrate that there is a high degree of variability amongst the data samples that is preventing adequate classification. This is particularly clear with the voice sample classes.

To address the between-video variability in the data, k-means clustering was explored. Using an unsupervised method would allow each individual video to be assessed without incorporating samples from other videos. Each 250 ms sample was converted to vector representation of the midterm features extracted by PyAudioAnalysis and standardized. Additionally, to aid classification, the samples were divided into 25 ms subsamples with 5 ms of overlap between each sample. The 25 ms samples consisted of short-term features extracted from PyAudioAnaylsis. Subsamples with an energy value

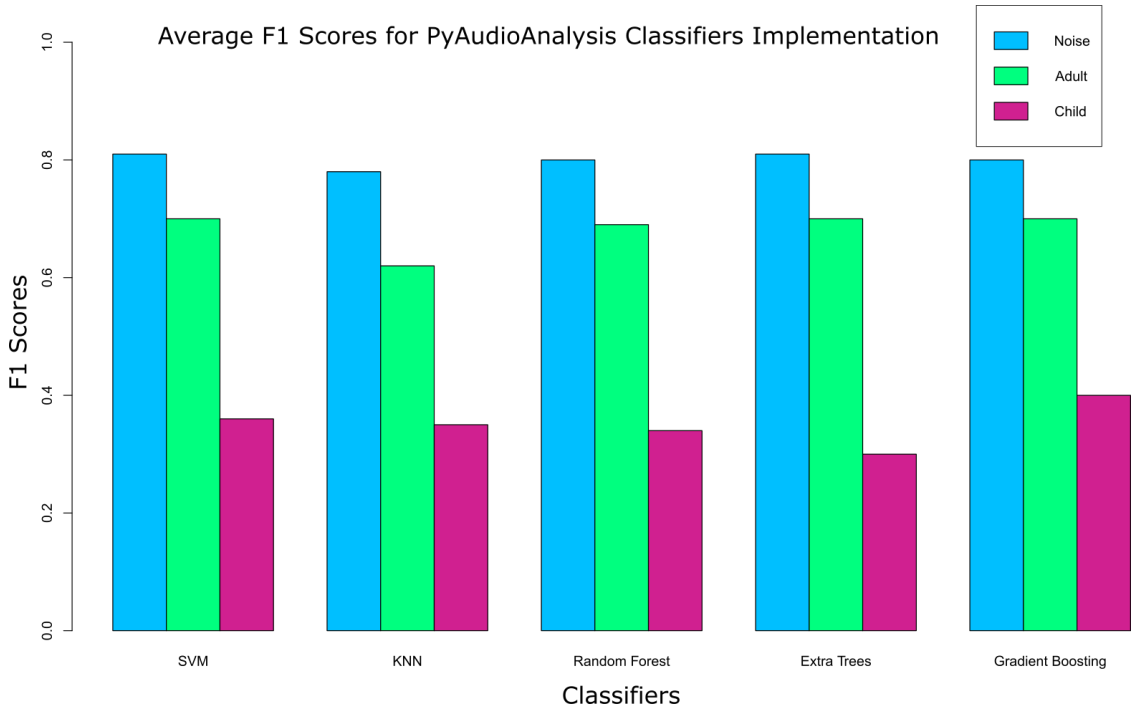


Figure 4.3: Average F1 scores from five classification algorithms in the PyAudioAnalysis library.

less than 1e-6 were discarded. The k-means algorithm was implemented using the sklearn python library (Pedregosa et al., 2011) with 10 maximum iterations.

The F1 scores from implementing k-means clustering are presented in Figure 4.4. These results varied between videos; however, performance was poorer than previous methods. Often, one cluster would dominate the data, accounting for the majority of samples. This was particularly true for the child and adult speech samples. A predominant issue with using clustering algorithms on this dataset is the level of data imbalance. The majority of the samples from each video are classified as noise, with a small minority of the samples coming from child utterances. In the cluster algorithm, this means that noise samples that have similar feature vectors to the speech samples will skew cluster centers, preventing the speech samples from created distinguishable groupings.

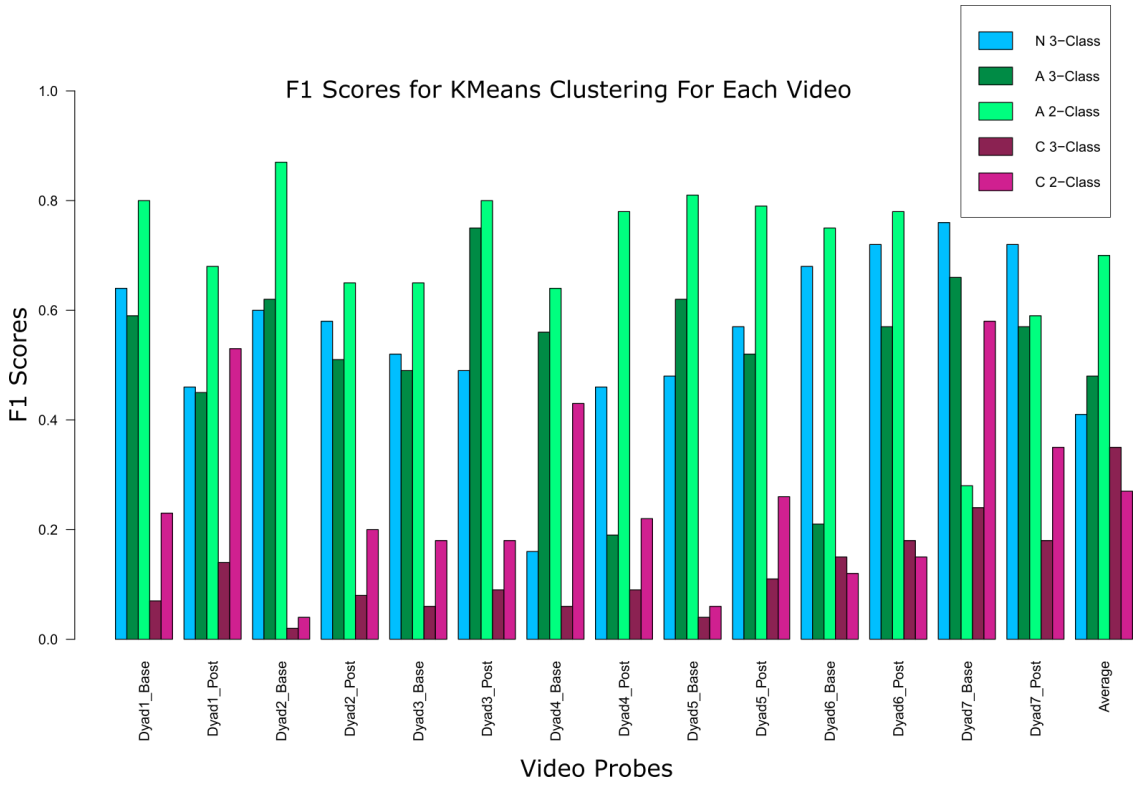


Figure 4.4: Child (C), adult (A), and noise (N) F1 scores from k-means clustering with PCA. Results are presented for using three classes and three clusters, and two classes with teo clusters. The two-class implementation excludes noise samples.

To account for the imbalance with noise samples, just the adult and child utterance samples were used in a two-cluster implementation. This shows improvement over the three-class classification; however, classification on child segments was still poor. This also could be due to data imbalance, as parent samples were more plentiful in the dataset. Child-directed speech patterns could also cause the adult speech samples to be similar to child samples, preventing effective cluster differentiation.

The final set of experiments revisited SVM implementation to explore approaching the VAD and speaker separation problems separately. To account for VAD, an SVM was trained using the noise samples as a class, and the combined adult and child speech samples as a second class. Similarly, speaker separation was accomplished by using child speech samples as a class, with the noise and adult samples as the second class. Both

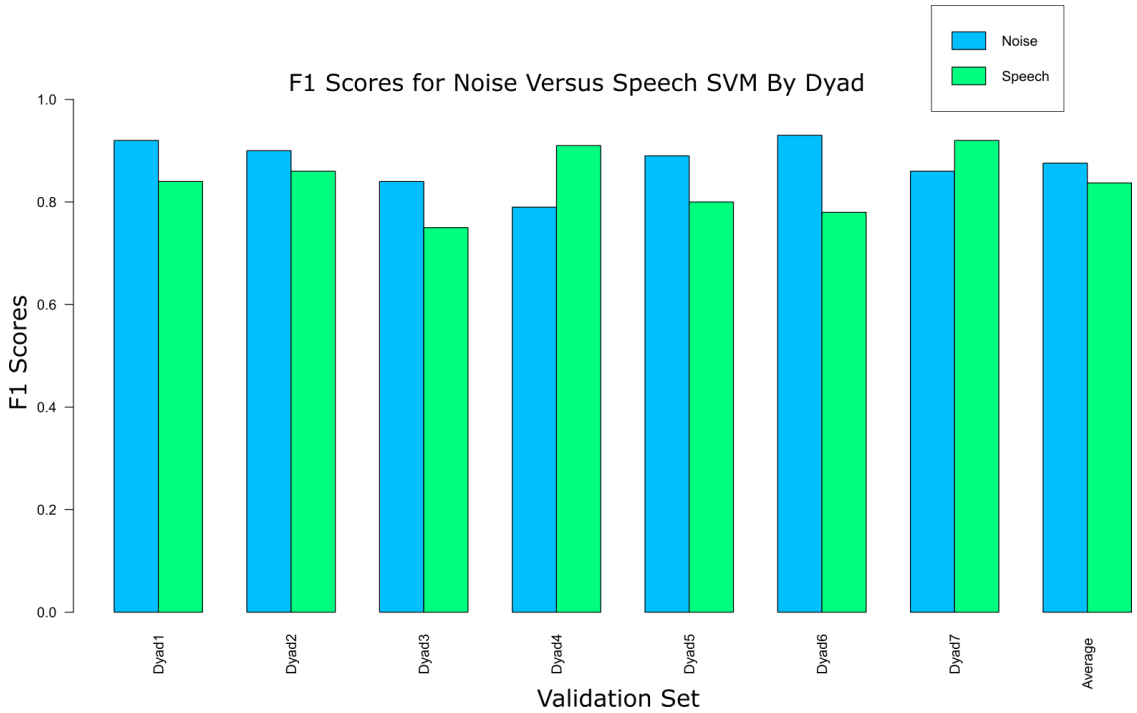


Figure 4.5: F1 scores for speech detection using a two-class SVM model.

SVM implementations used a C value of one and a RBF kernel. As with the PyAudioAnalysis experiments, the SVMs were trained using 12 of the 14 videos, using the remaining videos for validation, and the feature set consisted of the PyAudioAnalysis midterm extracted features. Data imbalance in the training set was addressed by undersampling the overrepresented class. The results are presented in Figure 4.5.

The separate VAD and speaker separation SVM implementations had a greater performance than the three-class classification techniques, particularly in distinguishing speech and noise samples. Classifying 250 ms segments on noise versus speech had an average F1 score of .85 over both classes across all seven validation sets Figure 4.5.

The average F1 score for speaker separation is lower than the VAD implementation, at .69; however, this is still higher than previous methods (Figure 4.6). In addition to testing 250 ms samples, the samples were divided into 100 ms subsamples with 25 ms overlap and used to train a separate SVM. Each 100 ms sample was processed through

PyAudioAnalysis to obtain the same feature set as previously noted. The goal of this was to determine if the subsamples were more diagnostic than the full sample. The F1 scores for the 100 ms samples was nearly identical to the 250 ms samples.

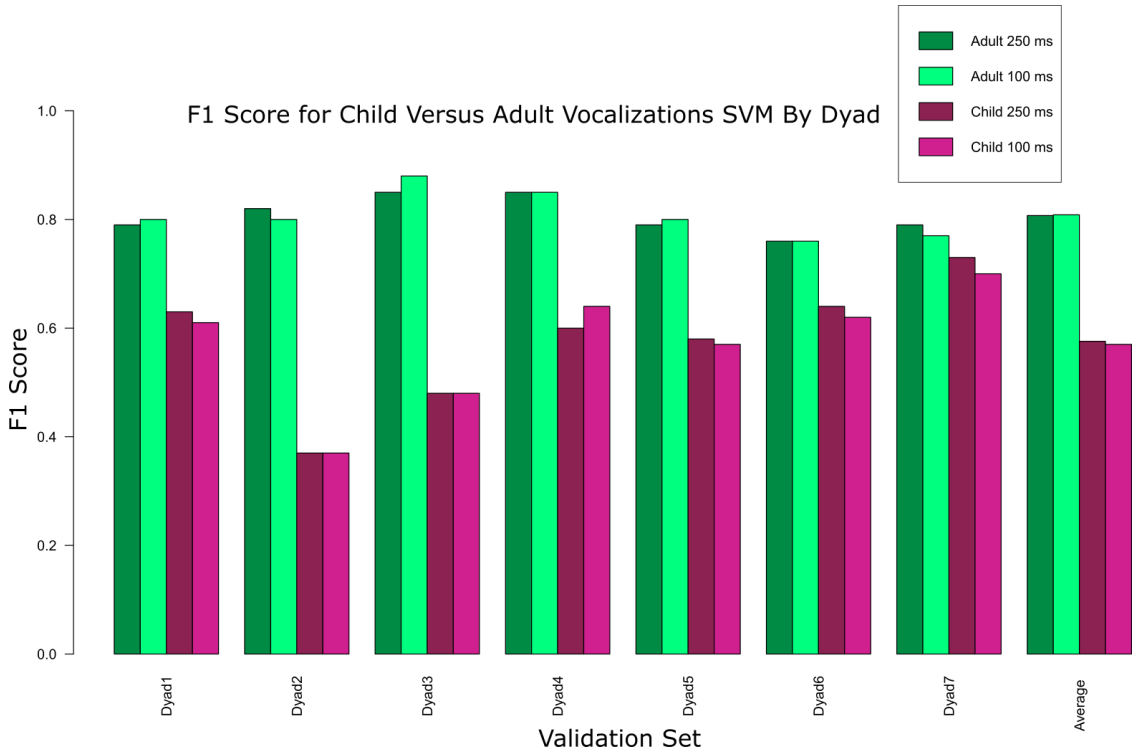


Figure 4.6: F1 scores for speaker classification using a 2-class SVM model. Results for 250 ms and 100 ms samples are shown.

The VAD and speaker separation SVM models were used to classify the audio in each of the 14 videos mimicking the intended implementation. The overall accuracy was 78%, with a range of 70 to 91% (Figure 4.7). Similar to the results presented in Figure 4.5, classifying noise samples had the highest accuracy at 87%. Noise samples are the highest represented class in the videos, leading this score to largely influence the overall accuracy. The speech accuracy was lower, averaging 65% for both classes. The Dyad 3 Post had the lowest accuracy for both the parent and the child at 42% and 46% respectively. This video had relatively low instances of vocalization for both individuals. The highest degree of error occurred by misclassifying speech as noise. Most of the

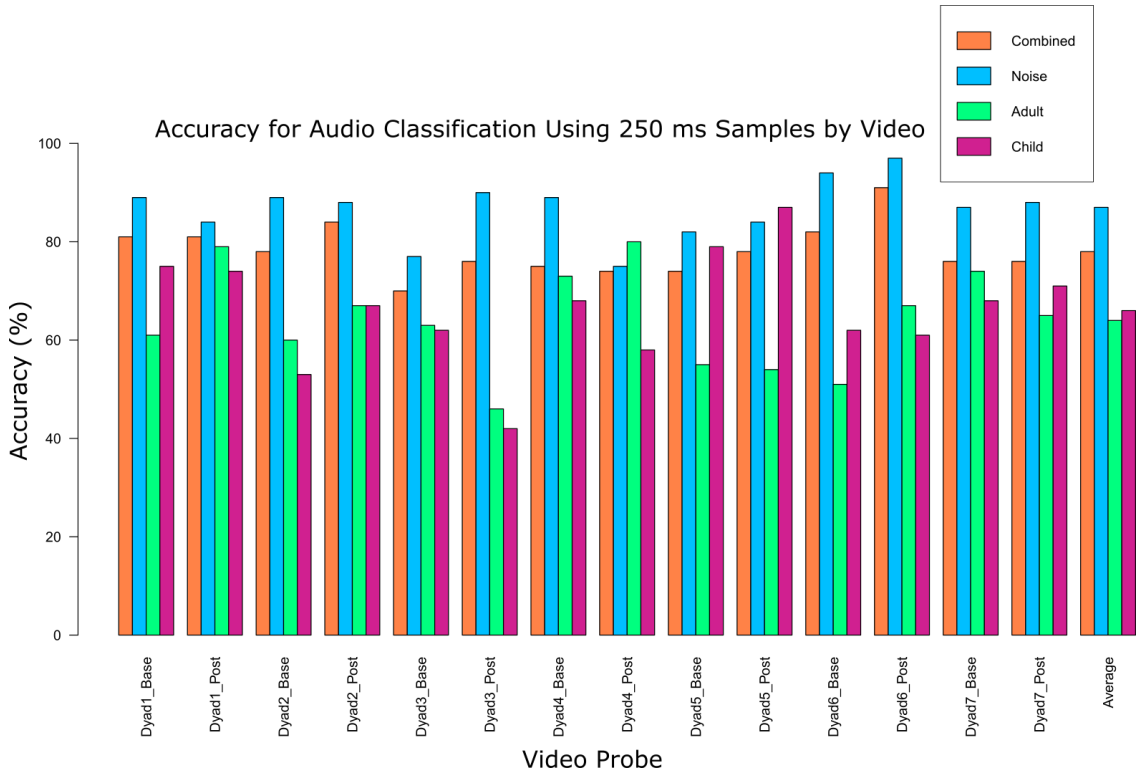


Figure 4.7: The classification accuracy for 250 ms noise and 250 ms speech segments in each of the PRT videos after running SVM models.

utterances made by the child in the video are attempts at the first phoneme of the prompted word. These attempts are generally short and clipped. This contrasts to the other children in the corpus that had longer vocalizations, even when they were only able to attempt a word. The parent in the video is drawing out words, pronouncing each syllable distinctly as an example for the child.

Each video was also evaluated for accuracy using the 100 ms speech classification model (Figure 4.8). The 100 ms samples were classified, then a label for the 250 ms segment was determined based on a voting scheme. This implementation had a similar overall accuracy of 79% compared to the 250 ms implementation. The average for both speech classes was slightly lower at 64%. The adult recognition improved over the 250 ms

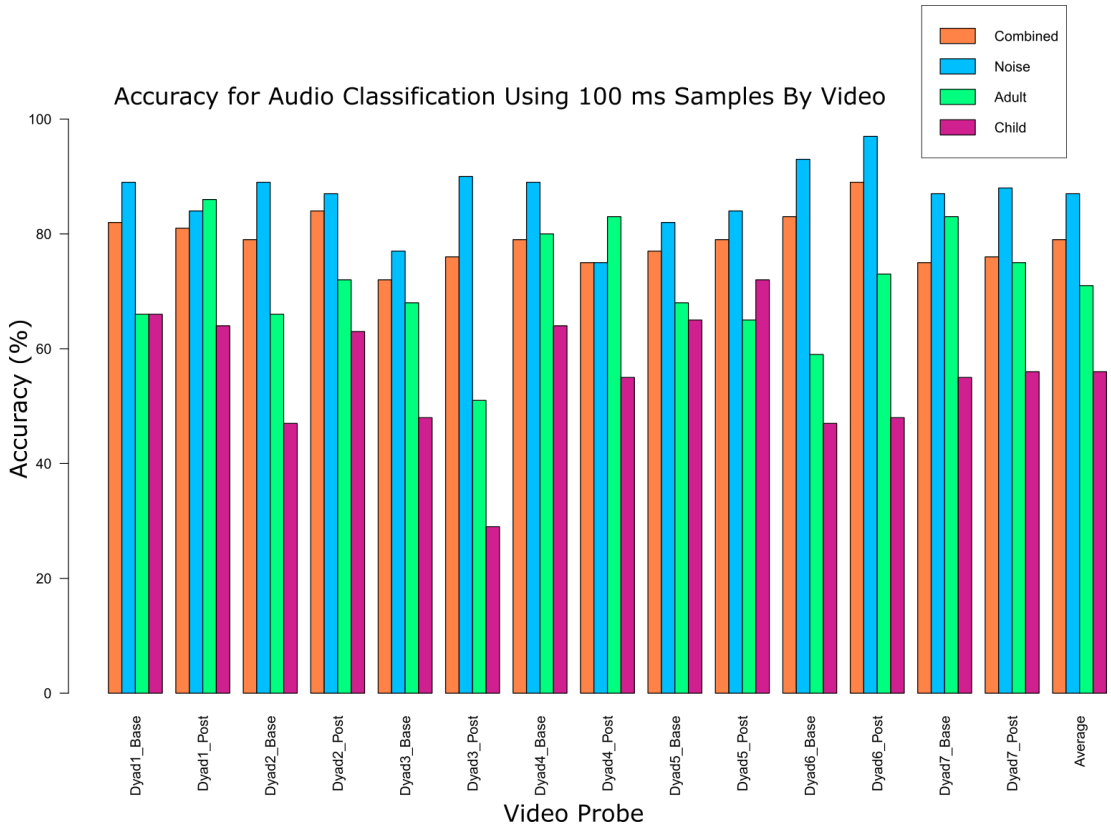


Figure 4.8: The classification accuracy for 250 ms noise and 100 ms speech segments in each of the PRT videos after running SVM models.

implementation; however, the child accuracy decreased. The increase in overall accuracy is due to the adult samples being more numerous than child samples in each video.

## 4.4 Discussion

When considering obvious differences between adult and child speech, pitch becomes one of the key components. As was shown in the pitch estimation analysis (Figure 4.2) and the results from the rule-based classifier, differentiation of pitch can be seen between sample classes. However, pitch alone could not be fully utilized to discern the vocal samples. As seen in the corpus and in other research studies, the most common composition for the parent-child dyad is an adult female with a child male, which have more similar vocal

frequency than may be present with other compositions. In addition to this, the adults in the videos have been shown to utilize child-directed speech, raising the intonation of their speech. This further limits the differences in frequency between the child and the adult. This necessitates exploring more features for speaker separation and the creation of classification models.

Evaluating the PRT audio corpus illustrates that a large degree of variability can be expected. Examining the participant's age range and communication skills accounts for much of the difficulty in creating a generalized solution with limited data. Child development rate is an important factor, with large differences between children at 24 months and 60 months. This is elevated when differences in development rate are factored in. These factors complicate the training of adequate models to encompass the dataset, making overfitting a large problem. This can particularly be a problem with deep learning algorithms using a small dataset. This led to the decision to focus on traditional machine learning implementations.

As PRT is implemented on a wide range of individuals of all ages and communication abilities, it is necessary to look for ways of addressing these large variations. This is illustrated by the three-class classification results presented in Figure 4.3. These results show moderate performance on distinguishing noise and adult samples, but a low performance on child sample classification. This is likely due to the underrepresentation of similar child data samples across the videos. The lowest average child F1 scores were seen in Dyad 2 and Dyad 3. The children in these videos exhibited few fully formed words, with very different patterns. The child in Dyad 2 is the youngest amongst the dataset. His vocalizations are largely akin to babble. The child in Dyad 3 communicated in short attempts at a specific word.

In this study, we examined unsupervised clustering to address variability between videos. The clustering algorithm allowed each video to be classified only on samples from

the same video. This eliminates model confusion based on sample variation in the same class. In terms of the PRT corpus, this means that the model was not trying to associate the limited vocal attempts from the child with less developed communication skills with the more articulated speech from other videos. This approach proved to be impractical for the PRT videos, largely due to data imbalance, along with between-class similarity. The results in Figure 4.4 show that the child samples, which are underrepresented in each video, are poorly differentiated from the noise or adult classes. This is still an issue after removing the noise samples to perform the classification only on the speech segments. The larger number of adult samples causes the adult class to have a greater influence on the clusters in the algorithm. This, along with the prevalence of outliers that are similar to child samples, could prevent the clusters from adequately distinguishing between samples.

Ultimately, the best results on the dataset were achieved by training separate classifiers for differentiating between noise and speech, and child vocalizations from adult speech. The VAD classification performed adequately across the dataset. This is congruent with the results from the three-class classifier results. This shows that much of the ambiguity in the data is in the speech samples.

Spot checking the full video classifications showed several trends in misidentified segments. For speech segments, adult samples labeled as child speech often contained low energy speech or have limited amounts of speech in the segment. This was most commonly seen at the end of a multi-segment vocal event where the trailing speech was presenting in a portion of the last labeled segment. The misclassification was also more prevalent if the trailing syllables of the word were elongated. When full vocal events spanning multiple labeled segments were misclassified, the adult speech often had more inflection and higher tone, which is typical of child-directed speech.

Child speech segments that were classified as adult vocalization often didn't consist of speech or attempted speech sounds. Most commonly the misclassified segments were

excited babbling or higher pitched vocal sounds. This is likely due to examples of exaggerated excitement present in the adult training set.

Misclassifying either adult or child vocal segments as noise was most commonly due to the segment containing sounds other than the vocalizations. This could also occur in segments where the vocalization was low energy. In noise segments misclassified as adults, sounds from the adult not associated with speech, such as coughing, were classified as adult speech. Interestingly, a toy's tinkling chime was consistently classified as adult speech in the Dyad 1 Base video. This is likely due to child-directed speech patterns utilized by the parent in the videos. Several of the women exaggerated excitement in their voices to engage the child. This reflects in higher pitched sounds that are not typical of conversational speech. Noises that were misclassified as child speech were either low energy or consisted of a brief sharp sound.

In the Dyad 1 Base video, the parent and child are playing with a toy that emits intelligible speech when being played with. These sounds were classified as noise. The toy's speech sounds are noticeably lower tone, resembling an adult male's speaking voice, than the child and parent vocalizations. Audible speech from a movie the parent and child are viewing in the Dyad 2 Post was classified in part as noise, as well as adult and child speech.

Future work regarding VAD and speaker separation in PRT videos should continue to focus on sample variability. This work utilized a feature set consisting of mid-term or short-term features extracted using PyAudioAnalysis. Additional work could be undertaken to explore which features most adequately capture the differences between adult speech and child vocalizations. Increasing the number of samples could also help account for the variability seen between participating dyads. Including more samples representative of each child's age and communication ability could aid classification. It may also be beneficial to use separate models or classes for different child ability or age

groups. Adding more data by using models pre-trained on other speech corpora could aid classification. Including more samples would also allow utilization of more data intensive algorithms, such as deep learning networks.

## 4.5 Conclusion

Classifying audio segments in PRT videos is a challenging problem due to the video capture techniques, atypical adult vocal patterns, and limited child vocal activity. Using a limited data corpus, adequate results were achieved by separating the VAD and speaker separation tasks between two SVM models. Incorporating more data samples and pre-trained models will likely produce greater accuracies by addressing the variability across sample videos due to the child's age and vocal acquity.

# Chapter 5

# MULTIMODAL PROCESSING AND CLASSIFYING OPPORTUNITIES TO RESPOND

Detecting whether or not a caregiver has created an 'opportunity to respond' is dependent on multimodal analysis of the interaction. An appropriate 'opportunity to respond' occurs when the caregiver has captured the attention of the child, and the caregiver has provided a clear instruction at the child's language level. For automated detection, the system needs to be able to determine when these two conditions are adequately met; however, for the current project, the analysis of the vocal activity of the caregiver will focus only on determining when vocalization has occurred. The language utilized by the caregiver, which would determine if the vocalization was an instruction and if the instruction was at the child's level, will not be evaluated.

Detecting attention and voice activity detection and speaker separation were examined in Chapters 3 and 4 respectively. Classifying the audio signal using common machine learning techniques was shown to be feasible, even when given conditions including child-directed speech patterns and non-vocal speech; however, the attention classification has significant room for improvement. Including the audio signal in the classification needs to be examined as a potential means for improving attention classification. The attention state of the child will likely be influenced by the environment. Adding the audio signal provides additional contextual information that may be useful for classification.

This chapter presents an evaluation on how the audio data and visual pose features extracted from the PRT videos could be utilized to improve the attention classification and detect when a sample segment from a PRT video is a candidate for being an 'opportunity to respond.' To do this, there are four primary questions that need to be addressed:

- How should the video and audio features be mapped?
- How should the data streams be fused for utilization in a classification model?
- How can confidence estimates be used to improve classification?
- How can the results of classification be aggregated to provide a meaningful resolution for clinician feedback?

After addressing these questions, it is apparent that the epistemic uncertainty within the dataset has a large effect on the performance of classification approaches. Results for detecting attention vary between validation sets depending on the activities depicted in the videos. In spite of the difficulty with detecting attention in the videos, there are promising results regarding using multimodal data to infer candidate samples that exhibit a correct opportunity to respond.

## 5.1 Related Research

Research into multimodal classification has focused on different methods for combining modalities for a singular classification task. These methods primarily consist of feature fusion, or early feature fusion, the combining of features from each medium to use as a classification for a single model, or decision fusion, also known as late feature fusion, which combines the output of separate models for all of the media to infer a new classification.

In addition to multimodal research, studies involving the use of confidence estimates pertaining to the probability that the model produced the correct predicted class are relevant to the current project. The PRT data is rife with both aleatoric and epistemic uncertainty. Using the confidence estimates produced by the classification models provides a means of addressing uncertainty.

### *5.1.1 Multimodal Classification*

Examples of relevant studies on multimodal classification on implementations are presented below. These primarily focus on the use of audio, lexical, or video data. This research covers speech recognition and analysis, affect and engagement detection, and human activity recognition.

### *5.1.2 Audio-Based*

Detecting humor based on acoustic and lexical modalities was presented by Bertero et al. (2016). Their approach followed a decision fusion methodology. Two CNN's were trained, one with lexical features and one with acoustic features, to predict when a punchline would occur in the TV show The Big Bang Theory. The decision merging utilized a SoftMax function to determine the final sample label. They validated the approach by comparing it against a conditional random field classification model, logistical regression, and an RNN implementation. Each model was evaluated using combinations of each modality. They conclude that the multimodal CNN approach produced the best results based on F1-score and accuracy.

Acoustic and lexical modalities have been used in a similar manner to detect deception in audio recordings. Mendels et al. (2017) extracted spectral and prosodic features for audio signals and used concatenation fusion with lexical features to train a hybrid LSTM-DNN network. Their network was tested against logistic regression and random forest implementation. They concluded that their approach had the best performance, and achieved the greatest record F1-score for the Columbia X – Cultural Deception corpus.

### *5.1.3 Visual-Based*

Using facial images and application logs, Bosch et al. (2015) evaluated affect detection. For the experiment, participants interacted with an educational software program. It collected data during usage that was stored in log files. These files were synchronized with a video being recorded of the participant's face during use of the application. They examined 14 different classification approaches, including SVM and Naïve Bayes algorithms. Both feature and decision fusion techniques were explored. They concluded that, on average, decision-based fusion approaches produced the greatest results based on AUC metrics. Metrics on face-only feature sets were nearly as high as the multimodal feature set. The authors state in their conclusion that the face-only feature set likely suffered from missing data due to occlusion from hand or head movements.

Similar to Bosch et al., Castellano et al. (2012) used video recording and log data from playing an educational game to detect affect. The researchers evaluated 17 combinations of features from the two modalities, presumably using concatenation feature fusion, with an SVM classifier. They validated their results using leave-one-subject-out cross-validation. Utilizing all of the features was reported at having the highest performance based on prediction accuracy.

Concatenation feature fusion was used on data regarding body language, facial features, and application logs for affect classification while using a tutoring program (D'mello and Graesser, 2010). Different combinations of the feature set were analyzed using linear discrimination with leave-one-out cross validation. Their results showed that multimodal feature sets outperformed unimodal sets.

Also evaluating affect during the use of a tutor system, Grafsgaard et al. (2014) extracted dialog and task actions from application use logs to be used along with facial and body pose data. Their approach used concatenation fusion and evaluated different combinations of feature sets. Linear regression and model averaging were used to predict

sample affect labels. Notably, compared to other research that used experts to evaluate and label affect, this publication relied on self-reported class labels. Their conclusion was that the multimodal feature set consisting of all of the collected modalities surpassed unimodality models.

RGB-D video and Lidar were used as modalities for human activity detection in a system designed for human-robot interaction. In their work, Moencks et al. (2019) created a new dataset under laboratory conditions based on common actions humans undertake at home or in the office. Their feature set consists of human pose data and distance metrics. They validated a classifier on the dataset by comparing multiple machine learning algorithms, with a DNN implementation having the highest reported accuracy.

### *5.1.4 Audio-Visual-Based*

Harwath and Glass (2017) used a novel approach for associating speech with images. The goal of the work is to train an unsupervised classifier to be able to learn associations of images and speech without the need for text transcriptions. Their approach utilized the spectrogram visualizing the waveform of speech in association with an image of a scene or object. Both the object image and the spectrogram are encoded to reduce dimensionality, then the inner product of the feature matrix is calculated. This calculated matrix is used to train a CNN. Validation of the model compared the results to a previous CNN implementation from the same authors (Harwath et al., 2016). The results were based on clustering accuracy.

An attention-based decision fusion method is discussed in Hori et al. (2017). The goal in this publication is to predict words in a sequence for automatic video descriptions. Image information related to object recognition, optical flow motion data, and audio information are utilized for the prediction tasks. Their framework utilizes different network layers for each modality and a decoder layer that performs the decision fusion.

This decoder layer has an additional activation layer that learns weights for the decision features for each modality. The authors compare this approach to a 'naïve' decision layer on two versions of the YouTube2Text dataset. Performance is reported as BLEU, METEOR, and CIDEr metrics, which are designed for evaluating natural language processing and machine translation tasks.

Audio-visual speech recognition was the focus of a publication by Mroueh et al. (2015). Their approach used late feature fusion to merge the final layers of separate DNN implementations for audio and visual data. The audio data was presented as a spectrograph image, while the visual data depicted the speaker's face between the nose and chin. They validate their model against unimodal implementations and two different methods for fusion techniques. Their best reported results based on phone error rate indicated that a multimodal technique using a SoftMax fusion layer had the greatest performance.

A recent work focused on using audio-visual data for determining affect and engagement of children with autism interacting with a robot. The research, conducted by Rudovic et al. (2018), used acoustic, video, and electrodermal data in addition to contextual information about the child participant to train a multimodal classification model. The audio, video, and electrodermal data was concatenated and used in an autoencoder to handle missing or noisy features. The contextual information was utilized as a means of providing additional parameters, such as age and gender, that could help personalize the model. The results were reported using interclass correlation, with the authors concluding the personalized network outperformed other implementations.

Using facial motions from videos in addition to acoustic data for speech separation was the subject of an article by Gabbay et al. (2018). For their approach, they used a video-to-speech neural network model to predict likely speech based on silent video clips. The speech predictions were then used to evaluate and filter speech predictions from noisy audio. They evaluated their method against audio only implementations and concluded

that their work had better performance. This was extended to predict speech from multiple speakers in Ephrat et al. (2018).

### *5.1.5 Calculating Confidence Estimations*

Obtaining confidence estimations in machine learning, particularly deep learning, is not a straightforward undertaking. SoftMax calculations that are often used as part of the classification process do not represent an accurate measurement for the model's confidence in its label selection (Guo et al., 2017). For SVM implementations, this has been addressed by using logistic regression on the distance a point is from the optimal hyperplane (Rüping, 2004). Similarly, the SVM implementation from Scikit-Learn used for the initial experiments for the framework utilizes Platt scaling regression techniques for its probability estimations (Pedregosa et al., 2011). Other methods have looked at temperature scaling the SoftMax values to better represent the network's confidence (Neumann et al., 2018) and cluster density models that rely on distance measurements to infer confidence (Ju et al., 2018; Subramanya et al., 2017).

Numerous methodologies have been proposed for calculating confidence estimations for deep learning networks. A common method is to use a Bayesian neural network to learn distributions for weights instead of discrete values (Kendall and Gal, 2017). Additionally, ensemble methodologies evaluating common loss functions have been explored (Lakshminarayanan et al., 2017).

### *5.1.6 Using Confidence in Classification Tasks*

Confidence estimation can be used to improve classification in numerous ways. In selective classification or hypothesis evaluation methods, multiple evaluation pathways are computed with the output being selected based on the highest confidence score. This can be done by using different classification methods, such as RNN and CNN

implementations (Zhao et al., 2017), or by setting a threshold confidence level and rejecting hypotheses below this value (Specia et al., 2009; Wang et al., 2011).

The confidence estimations can also be included as part of the network learning and optimization phase. DeVries and Taylor (2018) added a confidence loss based on a parallel layer in their network that penalized low confidence in a solution. Geifman et al. (2018) used confidence estimation to short-stop network training to prevent overfitting.

One example of the use of confidence estimates in improving human activity recognition examined pose estimation. Einfalt et al. (2018) explored improving pose detection in video data streams of swimmers. Their approach used confidence prediction in parallel branches representing past classifications, present data, and future predictions to estimate and track body points over time. This approach followed a similar procedure to multiple hypotheses testing where multiple paths were evaluated with low confidence paths being pruned.

### *5.1.7 Unreliable Labels*

Related to the idea of confidence estimation is the concept of unreliable data labels. This is particularly relevant given the subjective nature of discretely labeling human behavior. Zhao et al. (2011) and Sukhbaatar et al. (2014) addressed noise in labels by creating estimates of label probabilities. These probabilities are learned through training neural networks in a similar fashion to an autoencoder implementation. A similar approach was undertaken by Jindal et al. (2016) that utilized a deep network for clustering samples to infer proper labels. In a publication looking at sparse data, Li et al. (2017) used an autoencoder to evaluate EEG data samples regarding human task engagement levels. Also looking at similarity metrics, Bootkrajang and Kabán (2014) used cross-validation and labels with trusted samples to address label noise in regression tasks.

## 5.2 Methodology

Two research goals are examined in this chapter. The first goal was to determine how the audio data can be used to improve the classification accuracy for detecting attention. This was explored by looking at feature concatenation by combining the audio and visual pose data into a single vector to use as an import for training a classification model, and by decision fusion using the prediction and confidence estimates for separate audio and video models to train a final classification model.

The second goal was to create a classification model for detecting samples that could contain an opportunity to respond. As with the multimodal detection model, the opportunity to respond model used both feature and decision fusion methods. For feature fusion, the features from the audio and video datasets were combined to train a classifier directly on the binary classification task of determining an opportunity to respond. Using decision fusion, the results from an attention classification model trained on video data and a speaker separation classifier were combined to infer if an 'opportunity to respond' occurred.

The scikit-learn (Pedregosa et al., 2011) support vector machine (SVM) implementation was used for the classification tasks. The SVM was trained with a C value of 10, gamma value if 0.001, and radial basis function (RBF) kernel. These parameters were chosen because they provide the most accurate probability estimates for the class predictions. Using the parameters, along with changes to how the videos were sampled, caused slightly different performance metrics compared to those posted in Chapter 3. Due to processing times, the 30 ms feature sets models were created using scikit-learn's ensemble package. This was implemented using the same SVM parameters as the 250 ms models on fifty estimators.

In order to properly fuse the two data modalities, a common sample size needed to be determined. The video dataset was labeled for attention at one second, 30 frame increments, while the audio labelling occurred in 250 ms segments. Four methods for mapping the data were used in the experiments. First, the one second segments were divided into the individual frames, with each frame having the same label. The feature set for each frame consisted of the normalized data extracted using OpenPose (Cao et al., 2017). The audio data was processed on approximately 33 ms samples. The short-form representations of the audio features were extracted using PyAudioAnalysis (Giannakopoulos, 2015) as discussed in Chapter 4. Each sample was labeled based on the 250 ms sample it was extracted from.

The remaining three methods for mapping the data were based on 250 ms samples. The audio features were congruent for all three methods and directly reflect the labeling and feature extraction process presented in Chapter 4. For the visual feature extraction, four samples were taken from each labeled segment. Two of the methods involved dividing the segment into four subsegments, then selecting a single frame to represent the subsample. The frames with the most and least similarity to the other frames in the subsegment were selected. These will be referred to as centroids and outliers, respectively. The final sample set, referred to as the composite set, was created using the same methodology described in Chapter 3; however, only four subsamples were produced instead of six. These subsamples were created by providing an average based on the OpenPose data from eight adjacent frames. Frame overlap was allowed to account for dividing the 30 frames equally into four subsamples.

## 5.3 Results and Discussion

### *5.3.1 Feature Selection Comparison*

The classification performance metrics for each of the feature sets was compared in Table 5.1. The centroid, outlier, composite, and 30 ms sample sets were used to train an SVM model to detect attention using audio, video, and combined audio and video feature sets. Each of the three 250 ms sample sizes shared the same audio data. For these experiments, the audio data models were trained using the corresponding attention labels, not the speaker labels. Overall, the performance on each classification model did not vary significantly. The variation in validation accuracy between sets is only 0.05 between the lowest and highest averages. Similarly, F1 scores for each of the classes is similar between approaches. Unlike the validation accuracy, the training accuracy does show a significant difference between each set. The centroid and outlier audio and video combined models had training accuracies at 98%. This likely indicates that the model overfit the training data; however, this only had a marginal effect on the performance metrics. Conversely, the visual-only feature set models had training accuracies of approximately 70%, but had slightly better performance metrics. The audio-only model's metrics were slightly lower than when the visual features were included.

Looking more at the results for the Dyad 1 and Dyad 2 validation sets shows an interesting contrast (Table 5.2). The results for Dyad 1 illustrate that the audio features aid in the attention classification in the videos. The audio-only classification results are substantially higher than the visual or audio-visual models. The opposite is true for Dyad 2, where the audio-only classification was significantly lower than the video-only. In particular, the Dyad 1 Base video and the Dyad 2 Post video require greater examination.

The Dyad 1 Base video had an accuracy of 52% when using only the audio features, compared to 27% when using only the visual features. The video consists of the child

playing with three different toys: a tower game were the child places balls into holes and the ball slides to the bottom; a peg with star rings where the child can construct a tower; and, a mechanized Cookie Monster toy that speaks and eats cookies. Each of these toys have distinct sounds: the balls make a plastic-on-plastic banging noise, the stars emit chimes, and the Cookie Monster toy's motor and gears produce noise during its functions in addition to when it talks. These noises particularly occur when the child or parent is interacting with the toy, which indicates the child's attention is not on the parent. The parent also speaks often during both attentive and inattentive periods; however, generally when the parent speaks the toy noises are also present. Other than transitions between toys, there were not many instances of relative silence.

The child's positioning in the video could be problematic for extracting the visual features. For most of the video, the child has his back to the camera. This could make the OpenPose recognition less accurate, as well as having a profound effect on the ability to infer visual focus.

The post video for Dyad 2 was substantially different from the other videos in the dataset. The video depicts only one activity - the child sitting in the parent's lap while watching a video on a mobile phone. Because the parent is participating in watching the video, the activity is considered shared attention. This means the majority of the samples in the video shared the same class label. The video-only model correctly classified 61% of the samples. Having both individuals relatively still and facing the camera improved the pose estimation rates and made detecting visual attention easier. As the parent is holding the camera, this meant that the child's visual focus was always on her hands. Additionally, the proximity of the individuals likely aided the video classification as other shared attention activities, such as reading a book, would have a similar closeness.

Unlike many of the other videos in the dataset, Dyad 2 Post does not have strong audio cues. The parent speech is limited to single words to prompt the child to attempt to

say the movie's main character's name and praise after the attempts. The child vocalizations are limited to babble elicited after the parent's prompts. The audio from the movie is relatively lower energy and infrequent in the first half of the probe, but it does contain spoken adult speech. These periods of low energy are not common in other videos, that are generally filled with speech or toy noises. In particular, other instances of shared attention would have more adult speech, such as when reading a book. Additionally, a brief conversation between the parent and an additional out-of-frame adult occurs. This adult speech could cause misclassification of samples.

The dichotomy illustrated in the videos for Dyad 1 and Dyad 2 indicates that both audio and video data can be useful for detecting attention under different circumstances. Comparing the individual performance on the individual modalities to the models using both the video and audio feature sets shows that concatenating the modalities into a single input is not the best approach. In the case of both Dyad 1 and Dyad 2, the performance decreased, with the accuracy for both dyads being 40% with the multimodal model; however, as stated above, part of this performance decrease is likely due to overfitting. Evaluating these two examples provides an opportunity for looking at ways to incorporate both modalities to create a more general classification model or system of models. Analyzing the classification probabilities to determine prediction confidence estimates could provide a meaningful approach for utilizing both methodologies.

### *5.3.2 Classification Probabilities for Decision Fusion*

Examining only the correct cases, Figure 5.1 shows the class probabilities for the centroid sample set for all of the validation sets. This shows that there is little difference in the probability distribution between the visual only and the audio-visual feature sets. For each of these, the average probability is between 60-65%. For the shared class, the mean for the probability distribution was within the 2nd to 3rd interquartile range of the other two

Table 5.1
Comparison of SVM performance for audio and visual feature sets for classifying attention (attn), inattention (inattn), and shared attention.

| Frame Selection | Feature Set | Validation Accuracy (%) | Training Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Outlier | Audio-Visual | 0.39 | 0.98 | 0.26 | 0.31 | 0.47 |
| | Visual Only | 0.44 | 0.73 | 0.33 | 0.32 | 0.50 |
| Centroid | Audio-Video | 0.41 | 0.98 | 0.28 | 0.30 | 0.51 |
| | Visual Only | 0.43 | 0.73 | 0.33 | 0.30 | 0.54 |
| Composite | Audio-Video | 0.42 | 0.95 | 0.26 | 0.26 | 0.54 |
| | Visual Only | 0.41 | 0.61 | 0.30 | 0.30 | 0.50 |
| 250 ms | Audio Only | 0.38 | 0.84 | 0.26 | 0.26 | 0.50 |
| 30 ms | Audio-Visual | 0.41 | 0.76 | 0.26 | 0.32 | 0.51 |
| | Visual Only | 0.44 | 0.69 | 0.32 | 0.32 | 0.51 |
| | Audio Only | 0.38 | 0.52 | 0.28 | 0.26 | 0.46 |

Table 5.2
Detailed comparison of Dyad 1 and Dyad 2 feature configurations using centroid sample sets for classifying attention (attn), inattention (inattn), and shared attention.

| Feature Set | Validation Accuracy (%) | Training Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|
| Dyad 1 Audio-Visual | 0.40 | 0.98 | 0.09 | 0.39 | 0.64 |
| Dyad 1 Visual Only | 0.30 | 0.73 | 0.10 | 0.40 | 0.60 |
| Dyad 1 Audio Only | 0.49 | 0.83 | 0.16 | 0.33 | 0.66 |
| Dyad 2 Audio-Visual | 0.40 | 0.98 | 0.48 | 0.21 | 0.30 |
| Dyad 2 Visual Only | 0.55 | 0.72 | 0.67 | 0.35 | 0.34 |
| Dyad 2 Audio Only | 0.26 | 0.81 | 0.23 | 0.25 | 0.18 |

feature sets. For attention and inattention, the distribution mean fell well below the visual and audio-visual feature sets. This illustrates that, overall, the classification model was less confident when assigning classes using only the audio data. Incorrect classification probabilities that were comparable followed a similar distribution to correct predictions.

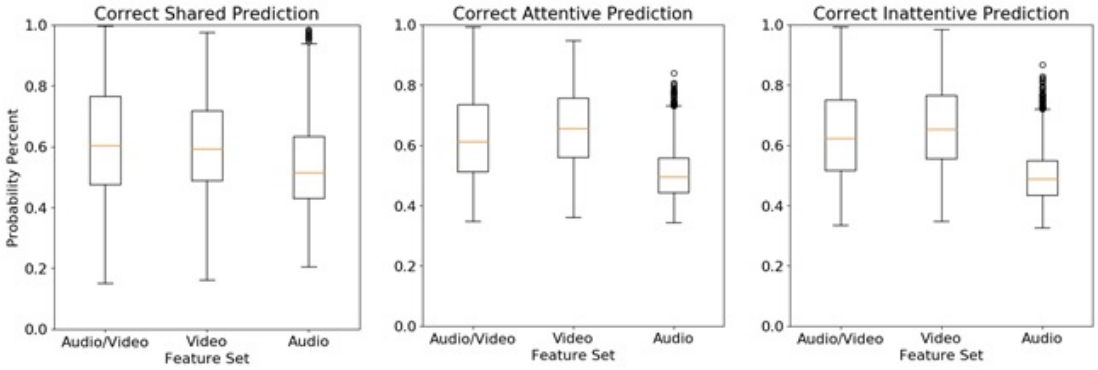


Figure 5.1: Box plots for the SVM probabilities across all dyads using the centroid sample set. The results are shown for correctly classified shared (left), attention (center), and inattention (right) samples.

The results presented above showed that the audio data was particularly diagnostic of the Dyad 1 validation set, while the visual-only feature set showed the best performance for the Dyad 2 validation set. This trend was not reflected in the examination of the class probabilities. Both the Dyad 1 (Figure 5.2) and Dyad 2 (Figure 5.3) showed similar distributions to the amalgamated results in Figure 5.1. These results suggest that simple decision fusion methods, such as using the highest probability classification, will be dominated by the video results.

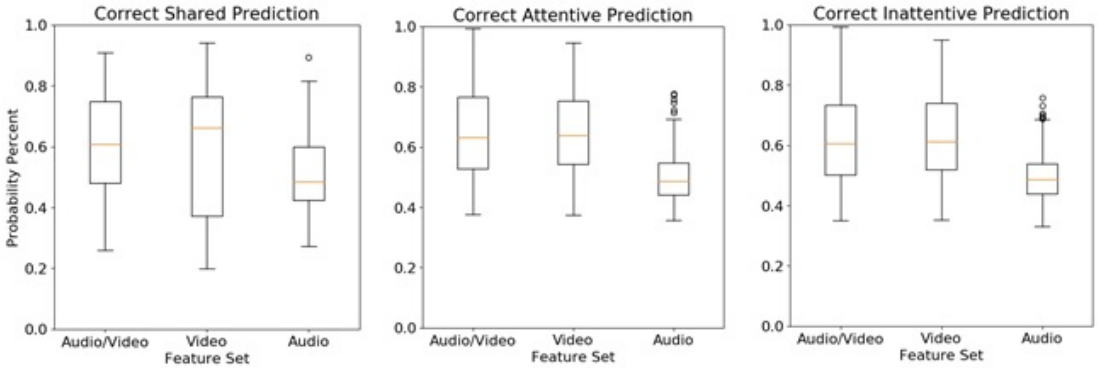


Figure 5.2: Box plots for the SVM probabilities for Dyad 1 using the centroid sample set. The results are shown for correctly classified shared (left), attention (center), and inattention (right) samples.

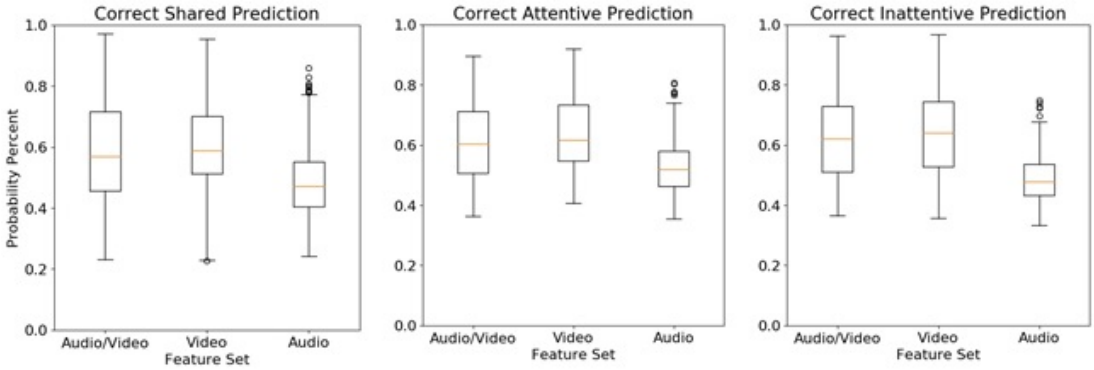


Figure 5.3: Box plots for the SVM probabilities across ofr Dyad 2 using the centroid sample set. The results are shown for correctly classified shared (left), attention (center), and inattention (right) samples.

To better understand using the multimodal fusion for attention classification, four methods were compared. First, the results for the audio-visual feature set presented in Table 5.1 represent early feature fusion, using a single input vector that represents the concatenation of the data from both modalities. Second, the SVM probabilities for the separate visual and audio sample sets are compared and the highest probability is selected as the prediction for the sample. Intuitively this would compensate for cases where one modality is more diagnostic than the other. Third, the probabilities for each class are summed between the audio and visual classification results and the greatest probability is selected for the prediction. This would address instances where a single class is not dominant for one of the modalities for a sample. Adding the second modality's probability estimate would allow for a greater distinction. The final method is to train a decision tree classifier on the probability estimates. Using a decision tree would aid in discerning a pattern in the probability's relation to the true class. A depth value of three was used for the decision tree. It achieved an average training accuracy of 74%.

Table 5.3
Comparison of fusion methods for audio and visual feature sets for classifying attention (attn), inattention (inattn), and shared attention. Values based on an average over all seven validation sets.

| Fusion Method | Validation Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|
| Feature Concatenation | 0.41 | 0.28 | 0.30 | 0.51 |
| Max. Probability Decision | 0.43 | 0.33 | 0.31 | 0.49 |
| Sum Probability Decision | 0.43 | 0.31 | 0.32 | 0.47 |
| Probability Decision Tree | 0.45 | 0.27 | 0.28 | 0.45 |

The fusion results are represented in Table 5.3. These results do not show a significant difference between each fusion methodology. The decision tree implementation had the highest overall accuracy, while feature concatenation had the

lowest accuracy. The two decision fusion methods using probabilities had similar results, and presented greater F1 scores for the shared and attentive classes. It is not surprising that these two methods produced similar results when considering the probability distributions.

Examining the decision fusion methodology of selecting the highest probability between separate audio and video classifiers illustrates that the video classifications are dominant. Table 5.4 presents the metrics for each of the validation sets. On average only 23% of the class labeling used the audio classifier prediction. Overall, selecting the audio classification was also less accurate, only being correct 38% of the time. The visual classifications were correct on average 44% of the time. The notable exceptions are Dyad 1 and Dyad 2, which remained congruent to the results discussed above, with Dyad 1 having a higher accuracy using audio data and Dyad 2 achieving better performance with visual data. Also interesting is that there is a low rate of agreement on correct predictions. On average, both models only predict a correct label on 36% of the samples.

Table 5.4
Comparison of results from assigning a class to each sample based on the highest probability between independent audio and visual classification models for each validation set.

| Val. Set | Predicted with Audio(Visual) (%) | Acc. (%) | Audio Pred. Correct (%) | Video Pred. Correct (%) | Predicted Audio When Visual Correct (%) | Predicted Visual When Audio Correct (%) | Both Correct (%) |
|---|---|---|---|---|---|---|---|
| Dyad 1 | 0.21(0.79) | 0.35 | 0.52 | 0.31 | 0.14 | 0.34 | 0.35 |
| Dyad 2 | 0.24(0.76) | 0.48 | 0.21 | 0.56 | 0.42 | 0.14 | 0.26 |
| Dyad 3 | 0.31(0.69) | 0.36 | 0.37 | 0.36 | 0.20 | 0.22 | 0.23 |
| Dyad 4 | 0.18(0.82) | 0.46 | 0.38 | 0.48 | 0.28 | 0.20 | 0.48 |
| Dyad 5 | 0.25(0.75) | 0.46 | 0.40 | 0.48 | 0.22 | 0.21 | 0.43 |
| Dyad 6 | 0.23(0.77) | 0.46 | 0.45 | 0.46 | 0.23 | 0.21 | 0.48 |
| Dyad 7 | 0.23(0.77) | 0.42 | 0.31 | 0.46 | 0.29 | 0.15 | 0.31 |
| Average | 0.23(0.77) | 0.43 | 0.38 | 0.44 | 0.26 | 0.21 | 0.36 |

Comparing results for each validation set between the decision tree method for decision fusion (Table 5.5) and feature fusion using concatenation (Table 5.6) shows that the accuracy increase is pronounced for three of the validation sets, while two sets showed little improvement, and the remaining sets had slightly lower accuracts. Dyads 4, 5 and 8 had a solid improvement using the decision tree, with Dyad 5 showing the greatest improvement with an accuracy of 55% compared to 43%. However, examining the F1 scores shows a decrease in the discernibility of shared class samples and a slight increase in the F1 score for inattentive class samples. This illustrates that the tree classifier is selecting inattentive samples at a greater rate than the feature concatenation method. The increased accuracy is a result of the data imbalance problem.

Table 5.5
Comparison of results for each validation set using a decision tree to perform decision fusion using SVM probabilities on centroid sample sets. Results are presented for overall validation and training accuracy, and F1 scores for shared, attentive (attn), and inattentive (inattn) classes.

| Validation Set | Validation Accuracy (%) | Training Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|
| Dyad 1 | 0.39 | 0.75 | 0.18 | 0.6 | 0.29 |
| Dyad 2 | 0.41 | 0.73 | 0.46 | 0.19 | 0.27 |
| Dyad 3 | 0.34 | 0.75 | 0.17 | 0.28 | 0.32 |
| Dyad 4 | 0.53 | 0.74 | 0.06 | 0.29 | 0.65 |
| Dyad 5 | 0.55 | 0.74 | 0.3 | 0.15 | 0.62 |
| Dyad 6 | 0.51 | 0.75 | 0.32 | 0.13 | 0.56 |
| Dyad 7 | 0.44 | 0.75 | 0.38 | 0.29 | 0.46 |
| Average | 0.45 | 0.74 | 0.27 | 0.28 | 0.45 |

### *5.3.3 Classification Probabilities for Sample Selection*

Label subjectivity and class imbalance are two inherent issues with classifying attention in the PRT videos. While label subjectivity has not been addressed, class imbalance has been approached by undersampling high volume classes to match the lowest represented class.

Table 5.6
Comparison of results for each validation set using concatenation to combine audio and visual features into an input vector for training an SVM model. Results are presented for overall validation and training accuracy, and F1 scores for shared, attentive (attn), and inattentive (inattn) classes.

| Validation Set | Validation Accuracy (%) | Training Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|
| Dyad 1 | 0.40 | 0.98 | 0.09 | 0.39 | 0.64 |
| Dyad 2 | 0.40 | 0.98 | 0.48 | 0.21 | 0.30 |
| Dyad 3 | 0.35 | 0.98 | 0.24 | 0.38 | 0.29 |
| Dyad 4 | 0.46 | 0.98 | 0.08 | 0.45 | 0.66 |
| Dyad 5 | 0.43 | 0.98 | 0.37 | 0.20 | 0.65 |
| Dyad 6 | 0.48 | 0.99 | 0.39 | 0.25 | 0.63 |
| Dyad 7 | 0.33 | 0.98 | 0.30 | 0.23 | 0.43 |
| Average | 0.41 | 0.98 | 0.28 | 0.30 | 0.51 |

The expectation is that utilizing the prediction probabilities for the training sets would offer a way of solving these sampling issues. Additionally, the intuition is that this approach will address the overfitting problem that was apparent in the audio-visual SVM models in Figure 5.1.

The results presented in Table 5.7 do not support the intuition for employing the prediction probabilities to perform a sample dropout. Two sample sizes, 4000 and 2000, were used to evaluate the methodology. These sample sizes reflect the minimum class representation in each training set (Table 5.8). Reducing the sample size did achieve the goal of reducing the training accuracy in order to address overfitting; however, this did not result in an improvement of training accuracy or average F1 Scores.

### *5.3.4 Opportunity to Respond Classification Evaluation*

Labeling the dataset for an opportunity to respond was based on combining the labels for attention and speaker separation. This is a binary classification problem with a positive label being attached to a sample where the attention state is either attentive or shared and

Table 5.7
Comparison of the average results for using the training probabilities to drop low confidence samples for training a second SVM. Values indicate validation and training accuracy, along with shared, attention (attn), and inattention(inattn) F1 scores.

| Feature Set | Max. Samples | Validation Accuracy (%) | Training Accuracy (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Audio-Visual | 4000 | 0.37 | 0.71 | 0.10 | 0.32 | 0.49 |
| | 2000 | 0.38 | 0.71 | 0.15 | 0.33 | 0.50 |
| Audio | 4000 | 0.38 | 0.71 | 0.15 | 0.33 | 0.50 |
| | 2000 | 0.39 | 0.62 | 0.22 | 0.30 | 0.50 |
| Visual | 4000 | 0.42 | 0.71 | 0.30 | 0.32 | 0.53 |
| | 2000 | 0.41 | 0.61 | 0.16 | 0.33 | 0.53 |

Table 5.8
Number of samples for each class (shared, attentive, inattentive) for each validation set.

| Validation Set | Shared | Attentive | Inattentive |
|---|---|---|---|
| Dyad 1 | 8577 | 4800 | 10755 |
| Dyad 2 | 6413 | 4989 | 11689 |
| Dyad 3 | 7004 | 4683 | 11285 |
| Dyad 4 | 8719 | 4459 | 10490 |
| Dyad 5 | 7545 | 5006 | 9846 |
| Dyad 6 | 7794 | 4867 | 10484 |
| Dyad 7 | 7120 | 4952 | 11195 |

the audio label is adult speech. To map the modalities, the one second segments labeled for attention are divided into four subsegments, retaining the original label, and associated with corresponding labelled audio segments. These only represent candidate samples for determining an opportunity to respond. At this point, the algorithm is identifying only if the parent has vocalized at a time when the child was attentive. This does not account for the natural language processing task of evaluating if the vocalization was a proper instruction. Table 5.9 displays the number of 'opportunities to respond' candidate segments identified in each validation set in the dataset.

Table 5.9
Number of ‘Opportunity to Respond’ (OTR) 250 ms centroid segments for each validation set in the dataset.

| Validation Set | OTR Seg | Total Seg | OTR Seg Time (sec) |
|---|---|---|---|
| Dyad 1 | 336 | 2982 | 84.00 |
| Dyad 2 | 642 | 4022 | 160.50 |
| Dyad 3 | 1043 | 4141 | 260.75 |
| Dyad 4 | 691 | 3445 | 172.75 |
| Dyad 5 | 524 | 4716 | 131.00 |
| Dyad 6 | 452 | 3968 | 113.00 |
| Dyad 7 | 834 | 3846 | 208.50 |

Three methods for classifying multimodal data were explored for detecting opportunity to respond candidates. First, separate classifiers for audio and video data were trained on speaker separation and attention tasks respectively. A sample was determined to be an ‘opportunity to respond’ candidate if the attention classifier predicted shared or attentive and the audio model predicted adult speech. The second method used the prediction probabilities from the same classification models as the first method to train a decision tree. The probabilities were used as the input with binary opportunity to respond labels to train the tree model. The final method was to use feature concatenation to train a single SVM classifier on the concatenated feature vector for audio and visual data. The results of these methods are presented in Table 5.10.

In comparing the classification results, it is seen that the decision fusion methods had a higher accuracy than the feature fusion method. Using the decision tree provided a slight increase in accuracy and F1 scores over the comparison method. These scores are influenced by the data imbalance. The classification models predict that a sample is false, and due to the majority of samples being false, has an inflated accuracy. This is shown in the disparity between the true and false F1 scores. The F1 scores for the true class are

Table 5.10
The table displays results for decision fusion using sample comparison and decision tree methods, and feature fusion.

| | Comp. Dec. Fusion | | | Dec. Tree Fusion | | | Feature Fusion | | |
|---|---|---|---|---|---|---|---|---|---|
| Validation Set | Accuracy | F1 True | F1 False | Accuracy | F1 True | F1 False | Accuracy | F1 True | F1 False |
| Dyad 1 | 0.83 | 0.40 | 0.90 | 0.85 | 0.38 | 0.92 | 0.73 | 0.74 | 0.74 |
| Dyad 2 | 0.82 | 0.48 | 0.89 | 0.83 | 0.47 | 0.90 | 0.74 | 0.76 | 0.79 |
| Dyad 3 | 0.75 | 0.45 | 0.84 | 0.77 | 0.56 | 0.85 | 0.68 | 0.70 | 0.71 |
| Dyad 4 | 0.67 | 0.35 | 0.78 | 0.65 | 0.3 | 0.75 | 0.62 | 0.53 | 0.59 |
| Dyad 5 | 0.84 | 0.44 | 0.91 | 0.85 | 0.37 | 0.91 | 0.72 | 0.74 | 0.75 |
| Dyad 6 | 0.90 | 0.44 | 0.94 | 0.91 | 0.63 | 0.95 | 0.72 | 0.76 | 0.83 |
| Dyad 7 | 0.77 | 0.52 | 0.85 | 0.78 | 0.48 | 0.86 | 0.71 | 0.69 | 0.68 |
| Average | 0.80 | 0.44 | 0.87 | 0.81 | 0.46 | 0.88 | 0.70 | 0.70 | 0.73 |

below 50% for both decision fusion methods. This illustrates the classifier could not adequately distinguish when true samples were present.

The feature fusion results did not produce the same accuracy as the decision fusion methods; however, the improvement in the F1 score for the true class predictions provokes more confidence in the model's learning power. The average F1 score for the true class in the feature fusion method was 70%, while the F1 score for false predictions was 73%. This shows that the classification model is not defaulting to false in a majority of cases, as it was with the decision fusion methods. This indicates that it has learned some features for distinguishing the two classes; however, the problem is still a challenge for the model.

The greater accuracy of the speaker separation models (as shown in Chapter 4) over the attention models used in decision fusion methods dominated the decision fusion classification for determining opportunity to respond candidates. This caused the samples that were false due to the audio being noise or child vocalization to be easy to detect. When the speech was identified as from an adult, the prediction was left to the less accurate attention classification label to determine a final class label, causing the low metrics for the true class. The improvement in the F1 scores for the true class for the feature fusion method over the decision fusion methods is likely due to the classification

model using the audio features to overcome some of the ambiguity in the visual data used for the attention classification. This indicates that the audio features may be useful in improving the attention classification.

The results for Dyad 4 were an outlier among the validation sets, having an accuracy score roughly 10 points lower on all three methods. This is likely due to difficulties with the audio classification. The audio in both the base and post video is relatively lower energy, due to the recording and the caregiver, the child's mother, speaking quietly. Despite speaking quietly, the caregiver is animated during the play interactions with her child, often making audible noises mimicking the toys and using child-directed speech patterns. Additionally, a toy being used in the post video emitted loud noises and elicited exaggerated excitement in the caregiver vocalizations. These factors may not be significantly represented among the videos for the other six dyads. Without similar samples in the training set, the model was not able to classify the Dyad 4 validation set at the same performance level as the other sets.

### *5.3.5 Execution Performance Comparison*

The runtime performance metrics are presented in Table 5.11. These metrics are based on a Windows 10 execution environment with an Intel i7 eight core 2.50 GHz CPU and 16 GB memory. The increased number of elements in the audio-visual combined feature sets had a large effect on performance times when training and validating a three-class model. Performance was improved when using two classes, as illustrated by the 'opportunity to respond' SVM results which used the same feature set as the audio-video SVM. Overall, validation times were not substantial. Performance could be improved by using parallel computing to process segments of the data simultaneously. These metrics only include post video processing classification. The data extraction using OpenPose takes a considerable amount of time. This extraction was executed on a Windows 7 machine with

a six core AMD 3.2GHz CPU, 16 GB RAM and NVidia GTX 970 GPU. The NVidia CUDA platform was used to run OpenPose using the GPU. This process took an average of four hours and 12 minutes to run a 10 minute video.

Table 5.11
Comparison of the average time for training and validation for the different classification methods and feature sets.

| Model Type | Feature Set | Training Time (h:mm:ss) | Validation Time (h:mm:ss) |
|---|---|---|---|
| SVM | Audio-Video | 0:25:04 | 0:00:23 |
| | Video | 0:03:45 | 0:00:09 |
| | Audio | 0:07:28 | 0:00:08 |
| | Opp. To Respond | 0:02:53 | 0:00:03 |
| SVM + Dropout SVM | Audio-Video | 0:51:31 | 0:00:27 |
| | Video | 0:09:29 | 0:00:11 |
| | Audio | 0:33:01 | 0:00:15 |
| SVM + Decision Tree | Audio and Video | 0:17:46 | 0:00:08 |
| | Opp. To Respond | 0:12:43 | 0:00:07 |

## 5.4 General Discussion

The experiments presented in this chapter further illustrate the epistemic uncertainty that challenges classification models on the PRT dataset. Classification performance varies across the different methodologies for each validation set. This is illustrated by the differences in attention classification based on modality as observed in the Dyad 1 and Dyad 2 sets, which responded more favorably to audio and visual data, respectively. This can also be seen in the lower performance on the binary multimodal classification of 'opportunity to respond' seen in Dyad 4. These differences between sets makes it difficult to implement a system that adequately addresses each scenario.

Using the prediction probabilities as confidence estimates also proved problematic. Examining the distribution patterns between audio, visual, and multimodal feature sets did not exhibit an exploitable behavior. The classification models exhibited high probabilities for predictions based on the ST feature set. This caused the visual results to largely overwhelm the predictions based on the audio feature set during decision fusion.

Although detecting attention remains a challenging issue, more favorable results were obtained for detecting opportunity to respond candidate samples. Using a feature fusion model produced an adequate classification that exhibited more favorable F1 scores than the attention models. Being able to reliably detect 'opportunity to respond' candidates could be used to identity video segments that would be of interest to clinicians during PRT fidelity scoring and performance evaluation.

## 5.5 Conclusion

Detecting the child's attention state in the PRT videos remains a challenging problem. Adding audio data in addition to pose information extracted from video frames provided more information; however, it did not reliably address the epistemic uncertainty inherent in the problem. This is largely due to the variation in activities the participants are engaged in in the videos, along with the unstructured recording environment. Decision fusion, feature fusion and using prediction probabilities as confidence estimates methods were explored to combat the uncertainty, but no single method provided an improvement across all validation sets. Including more data would be useful in addressing this issue as the models would have a greater pool of samples to draw from to determine similarities.

The multimodal problem of detecting an 'opportunity to respond' produced more adequate results than attention classification methods. Of the three methodologies that were examined, the feature concatenation exhibited the greatest classification prowess. Although the feature concatenation method had a lower accuracy than the decision fusion

methods, the greater F1 score for the positive class reflects a greater capacity to predict the true label. This is especially important given the class imbalance in the validation sets.

# Chapter 6

## PERSON-CENTERED CLASSIFICATION MODELS

A person-centered approach is an important consideration when designing a feedback system for PRT. Implementation of ABA is itself a person-centered approach aimed at adapting treatments to the individual recipient (Baer et al., 1968). This is particularly important in PRT where the recipient drives the activities and selects the objects he or she is motivated by in that instance. This is inherently individualistic and a source of variation that needs to be considered in the design as these factors contribute to epistemic uncertainty when training classification models. The variation inherent with the parent-child dyad exists alongside other key variations regarding context, demographics, and skill levels. Accounting for these variabilities in system design would be difficult, therefore it is important to look for ways to generalize behavior and adapt the system to each dyad.

In PRT involving parents and their children, the parent is expected to observe his or her child to ascertain an object or activity they are interested in, in that moment. After identifying this, the parent is expected to interject him or herself in the activity and facilitate learning by prompting the use of maintenance and target skills. Imagining this scenario playing out for a myriad of individuals quickly shows the amount of variation that will likely occur. This variation could be related to contextual and environmental settings, activities involving the participants, characteristics and behaviors of the individuals, and variance in recording devices.

The context and implementation environment where the session will take place will be different for each dyad, whether this takes place at home, school, or in a clinician's office. These environments will provide different opportunities for activities or interactions, have varying levels of distraction, and constitute different factors that could

influence the data, resulting in challenges for machine learning classification. Although these environments vary between different dyads, it is likely that a single dyad will use environments in multiple sessions.

Variation can also be expected based on the participant's level of known PRT. Observing differences between baseline and post-training video probes illustrates the difference the treatment makes in the behaviors of the participants. Typically, there are more instances of child-attentive or shared attention states and more child vocalizations, and less examples of adult speech in post video probes. Adapting to progress could help make the application more robust and useful to the user.

The activities that the dyad participates in in the videos will vary. Unlike contrived ABA methodologies, such as discrete trial training, there is no predetermined activity in PRT sessions. As stated above, the child selects the activity based on what appeals to them at that moment in time. This leads the activities that can be depicted in the videos to be a function of the recipients' preferences, mood, and the activities afforded to them in the environment. As with environments, it is likely the same activities would be performed in multiple sessions.

The age, gender, and ethnicity of both the child and parent will vary between dyads. Variations in gender and ethnicity will influence classification tasks. Ethnicity could affect classification models if the individuals exhibit accents or mannerism that were not adequately represented in the models' training samples. Similarly, racial characteristics, such as skin color or cranial morphology, have been shown to affect recognition tasks in computer vision. The age of the child is an important variation consideration for the proposed system. Child speech recognition is known to be challenging, especially in younger children (Lee et al., 1999; Potamianos and Narayanan, 1998). This is due to relatively slower speech patterns compared to adults, greater frequency of mispronunciations, and variations in pitch and volume. Additionally, the age of the child

will likely influence the interactions in the videos. The activities preferred by younger children will likely be different than those of older children. Speculatively, how the parent treats the child will also be associated with the child's age. The parents are more likely to use child-directed speech and exaggerated exclamations with younger children. As the child ages, the parent will be more likely to use speech patterns that reflect adult exchanges.

The communication skills of the children will vary across dyads as well as over time. Children with autism can exhibit a varying degree of verbal communication skills ranging from non-verbal or singular word usage to phrases and sentences. It is important for the system to be able to evaluate the entire spectrum of this speech. Related to this, the tasks being utilized to develop the child's skills will differ between individuals. During the intervention, the parent will need to identify maintenance tasks and target tasks for his or her child. Maintenance tasks are intended to be skills the child has developed but needs to continue practicing. These also serve to provide easily achievable goals that help limit frustrations by fostering a sense of accomplishment and providing access to praise and motivators. The target skill is intended to be in the child's zone of proximal development and represent a challenge that promotes educational growth.

Variation in the dyad's access to technology could impact their utilization of the framework. The framework is intended to work with affordable technologies that are readily available. This is intended to reduce the cost and make implementation easier and more accessible for the parents. Using different recording devices could affect the results of the classification system if the video quality is low. Also, the skills of the person using the device to record the session could have an impact. The operator's skill will influence how well the participants are kept in frame and how stable the device is while filming.

Examining the dataset that was used in the foundational research for the framework, described in Chapters 2 and 3, illustrates these concerns about variation. Although the

videos in the dataset depict a similar environment, a training room at an autism resource center, the activities the dyad participates in varies. Often, these activities are depicted in only one video. Watching a movie on a cell phone, and spinning in an office chair, are two examples. Attempting to classify these activities solely on the training samples in the dataset would be unsuccessful.

The dyads in the videos are comprised of an adult female and a child male. This representation fails to account for inclusion of the genders in each role. Additionally, all the individuals in the videos have light skin tones. This could fail to account for challenges regarding the computer vision portion of the project regarding the recognition of individuals with darker skin tones. The age of the children in the videos ranges from 24 to 60 months. This is a considerable gap when considering child development. This is reflected in a diverse range of vocal communication abilities ranging from phoneme-only vocalizations to full sentence speech. Accounting for all these variations solely using traditional classification techniques would require extensive amounts of data and is likely not feasible.

The current attempt to address these variations was to look at generalization to create a base model. As presented in Chapter 3, OpenPose (Cao et al., 2017) was used for performing the computer vision tasks in the framework. It is assumed that this application could adequately extract the individuals in the scene from the environment, along with accounting for variations in the individual's physical appearance. The output of OpenPose is a set of Euclidean coordinate values indicating where in an image frame-specific landmark body points are for each individual. These body points were used to calculate additional information including the likely focus of the individual's gaze and the relationship between the parent and child's hands in the frame. The purpose of calculating this additional data is to extract generalizable characteristics of attention from the activity and look for common cues such as eye contact, reaching for an object, or stillness. This is

based on PRT literature (Koegel, 1988; Suhrheinrich et al., 2011); however, this does not address individual behaviors that indicate attention. People, particularly people with autism, exhibit different cues to indicate attention. It is not uncommon for a child with autism to look away from an individual or avoid eye contact while still being attentive.

Often described not only as a training methodology but a way of life, PRT is expected to be undertaken over a long period of time, adapting as the participants grow and their needs change. Likewise, the expectation is that a feedback system would be utilized frequently. The frequency of use along with maintaining the relationship with the clinicians affords the opportunity to provide additional information to the system that could be used to personalize classification models. Intuitively, by creating a mechanism for personalizing the feedback, framework problems with aleatoric and epistemic uncertainty can be addressed over time.

Addressing aleatoric uncertainty could be undertaken by providing feedback on the recording device, environment, and other issues that would affect data acquisition. This would aid the caregiver in creating a more effective recording context for PRT sessions, allowing the system more favorable conditions for performing classification tasks. Epistemic uncertainty could be addressed by adding additional information to the system that could be utilized to refine classification models or select diagnostic pathways. This could include providing information regarding the child's age or language ability to the system, along with introducing new labeled data for fine-tuning classification models.

This chapter examines how a person-centered approach could be used to address epistemic uncertainty in the classification models. This will be undertaken by examining several key research questions. First, the role of providing easily obtainable *a priori* information, particularly the age and exhibited verbal communication level, was examined to determine how this can affect attention classification and speaker separation. Second, providing additional labeled data samples was explored. The goal of this was to determine

if providing labeled samples of the dyad will have an effect on classification tasks and how many samples would be needed to influence the classifier. Along with this, the effect of mislabeled samples needed to be addressed. Finally, a framework for presenting the data to the clinician for labeling was theorized. This examined the interface for clinicians along with the methodology for selecting samples to label.

## 6.1 Related Research

Examining the areas where variation occurs in PRT helps illustrate how a person-centered approach could be incorporated to enhance the framework. This could consist of using both automated and human-in-the-loop methodologies to foster co-adaptation between the system and the user. Human-in-the-loop adaptation could focus on facilitating personalization of the system to the individual and addressing activity and behavior variations from the videos. Automated methodologies could provide cues to users to modify use of the system, addressing environmental and technological variations. In addition, the parent-child dyad, clinicians, and PRT are important parts of the systems that would be influenced by a person-centered approach. The information supplied by the clinician as part of the human-in-the-loop paradigm and the data inferred through automated processes can be utilized to retrain the classification models to provide more personalization.

### *6.1.1 Automated Adaptation*

Part of the feedback the system provides the user could be used to drive co-adaptation, particularly to address creating a more favorable climate for data extraction. Discussions with clinicians revealed that parents often do not conduct PRT in a conducive environment. This usually pertains to a high level of distraction, including other people in the room, cluttered play areas, or electronic media running in the background.

Automatically detecting these issues and informing the user would provide an opportunity for her or him to take corrective actions. Similarly, issues regarding device records may be addressed the same way, leading to more optimal recordings with less occlusion or audio noise.

Automatically adapting classifiers has also been researched. Garcia et al. (2018) improved human activity classification by automatically fine tuning hyperparameters, primarily window size and sample overlap, in an ensemble model. Charles et al. (2016) used confidence estimates to include unlabeled samples of individuals to personalize and improve human pose classifications. A similar approach was used by Kato et al. (2018) in multi-person pose estimation.

Although not discussed in the early phase of the project, assessment and adaptation could be relevant person-centered features driven by automated voice data collection. One example that would address an identified need in the PRT implementation is the evaluation of maintenance versus target skills for the child. In interviews with clinicians and parents, both parties stated that adapting PRT to new skills was often difficult. The system could aid this by evaluating the child response rates and pronunciation on specific words to determine words that have been mastered, need to be reinforced as a maintenance task, or are in an area of proximal development and should be a target task. Much like the assessment in the Adaptive Training Assistant (ATA) (Tadayon et al., 2018; Venkateswara et al., 2018), this assessment would be handled implicitly by the system. Unlike the ATA that augments the user experience automatically, the assessment would be used to explicitly inform the user and aid them in developing a personalized training plan for the child. Adherence to this plan could then be further assessed by the system. The system could further leverage person-centered computing to incorporate information on toys and activities favored by the child to provide example instructions that address target and maintenance skills.

### *6.1.2 Human-in-the-Loop*

Human-in-the-loop systems integrate autonomous computations with human operators to accomplish a shared goal. The roles of the human and autonomous agents in the system differ based on the problem being addressed and the approach that is undertaken. The role of the human in these systems can be categorized as being the primary actor, an expert discriminator, or a tutor.

Technology provides a means for providing additional information to human operators to make tasks safer and more efficient. Utilizing sensors and data processing techniques affords a means of cueing the human to potentially important situations. In these types of systems, autonomous agents and human actors collaborate to achieve a goal. A common example is autopilot systems for airplanes and semi-autonomous vehicles (Gruyer et al., 2017). The autopilot systems analyze the environment and contextual information to provide greater insight into a situation to inform the human operator. As technology increases, the amount of autonomous actions and support becomes more sophisticated. Current human-in-the-loop systems have begun to analyze the condition of the human operator in addition to circumstantial awareness. This emphasizes human considerations including comfort, strain, and fatigue to provide a more optimal experience (Chiang et al., 2010; Feng et al., 2016).

For difficult or critical problems and operations, a fully autonomous system may not produce sufficiently reliable results. Addressing these problems can utilize autonomous systems to a limited extent; however, human experts, individuals, or crowd-sourcing (Li, 2017), are required to make a final determination. In these systems, the human acts as a final discriminator. This is common in the health and medical domain where problems can lack sufficient data, there are large sample imbalances between classes, or crucial events are rare (Holzinger, 2016). This approach can also be useful for more subjective classification tasks, such as bias detection (Jong et al., 2018).

Human operators can be used in a system to aid in training agents to perform autonomous tasks. This is a common approach for teaching robot systems motor controls (Peternel et al., 2015, 2018; Suomalainen and Kyrki, 2017). For these systems, the human assumes the role of tutor, demonstrating and correcting behavior. The human is providing additional data to the system that can then be utilized for refining autonomous decision making.

Maintaining the relationship between the clinician and the parent is a pivotal part of the proposed framework. Largely, this has been done to promote efficacy in the parent, provide feedback on the minutiae of the interaction not feasible for automated assessment, and promote social pressure for continued compliance. Keeping the clinician as part of the system provides the opportunity to have an expert label additional data samples, as well as record contextual information about the dyad, particularly regarding the child's vocal performance. This allows for the clinician to act as a tutor for the autonomous classification process, providing additional expert information that can be used to refine models.

Evaluating the confidence values during the classification process would allow the system to identify sequences where the generalized models could not adequately assign a label. These sequences could then be provided to one or more clinicians for labeling and used to retrain the models to personalize the classification task to the users. This would allow the models to gradually be improved as the dyad utilized the system. This likely would show significant improvements when the child and parent participate in similar activities across multiple videos. Incorporating this approach for vocalizations would help adjust generalized models for voice activity detection and speaker separation. This would address the challenges of child-directed speech patterns for the parents, and provide greater accuracy at detecting child vocalizations.

The clinician's expert knowledge could also be used to adapt the framework by providing additional contextual information, particularly when the parent-child dyad has limited use on the system. The clinician could initialize the system to account for parameters, such as the child's vocal ability, to configure the system to favor distinct classification models. One of the most important aspects of having the clinician involved in the system is to identify plateaus in the child's performance, and make recommendations to the parent on how these can be addressed. Additionally, the clinician's initial evaluation of the adult's familiarity with PRT could be useful contextual information for adapting the system.

The research provided by Xu et al. (2016) provides a scaffolding for how the co-adaptation could be implemented in the proposed framework. In their work, Xu et al. examined data from wearable sensors to classify and assess human activity. They utilized contextual information along with classification confidence levels to select appropriate models for detecting behavior from a network of ensemble machine learning algorithms. This allowed for the system to adapt to the individual and assess them effectively as they improve in the activity. Like this implementation, the proposed framework could utilize the contextual information to favor specific classification models that would better suit the individual. It could also evaluate the confidence levels in order to determine when it would be appropriate to obtain the aid of a domain expert to incorporate additional labeled samples.

### *6.1.3 Classification Fine-Tuning*

Deep learning algorithms are powerful tools for artificial intelligence; however, these approaches require a plethora of labeled data samples to achieve optimal results. For many domains and specific classification tasks, there is insufficient data for training these types of classification models. Fine-tuning is a methodology using transform learning to

leverage classification knowledge on one set of data to be used on a separate set of data. In addition to addressing a lack of data, using pre-trained networks can reduce the training time and reduce the likelihood of the network overfitting (Campos et al., 2017). Fine-tuning is undertaken by training a deep learning network on a substantial dataset, commonly ImageNet (Deng et al., 2009), then adding additional layers trained on the target dataset. It is also common to use a pre-trained model such as AlexNet (Krizhevsky et al., 2012) or Inception (Szegedy et al., 2015).

Fine-tuning classification exhibits the best performance when the initial training data has a similar structure to the fine-tuning data (Chu et al., 2016). The key concept is that the pre-trained network has learned to distinguish important features in the data. Having learned these features, adding the new data allows the next work to use the pre-established weights to extract features and relate them to the new target labels. Better results can be obtained by using only the subsets of larger datasets that best relate to the new target data (Ge and Yu, 2017).

Different methodologies can be used for fine-tuning networks. An alternative approach is to expand the nodes in the final classification layer of the network instead of deepening the network with a new layer specific to the target task (Wang et al., 2017). Using multi-task training has also been explored, where the initial training data and target training data are jointly presented to the network during the training processing (Ge and Yu, 2017).

Literature surrounding fine-tuning has primarily focused on visual data, with object classification and scene description tasks being common targets. It has an important application to the medical field (Kumar et al., 2016; Tajbakhsh et al., 2016) where labeled data on a specific task can be parsed. Research surrounding fine-tuning image-based networks with medical imaging has been favorable, showing that the networks can utilize feature information extracted from general visual data. Other novel applications include

specific classification, such as plant taxonomy (Reyes et al., 2015) and sentiment analysis (Campos et al., 2017).

For the PRT data, there are two types of fine-tuning that have been employed. First, as was detailed in Chapter 3, a pre-trained AlexNet model was fine-tuned with images from PRT videos to classify the child's attention state. Second, using fine-tuning to personalize classification models to a specific dyad was explored. This implementation is different from the common research applications. For personalization, the goal is not to perform transfer learning, it is to refine the classification domain to focus on an individual dyad.

## 6.2 Methodology

The research aim of this aspect of the project is to examine ways the clinicians could provide information that could be used to improve attention classification accuracy. There are three methods that need to be explored. This will be undertaken using the visual pose feature set using centroid samples as presented in Chapter 3 and the concatenated multimodal audio-visual feature set described in Chapter 5. As with previous experiments, data imbalance is handled by undersampling.

First, the clinician could provide *a priori* information about the child's age or demonstrated vocal communication ability that could be used to select a classification track that uses models trained on data from children of a similar demographic. This method has been demonstrated on similar tasks by Rudovic et al. (2018). For this chapter, this will be simulated by dividing the dataset into two groups based on the language skills apparent in the videos. This has been summarized in Table 4.1 in Chapter 4. Dyads 1, 2, 3, and 6 were included in the first group. These children were younger and exhibited one single word response or vocal attempt. Dyads 4, 5, and 7 were included in the second

group. Separate SVM classifiers were trained for each group using the SVC package in scikit-learn (Pedregosa et al., 2011).

A second way of using prior knowledge to segment the training data is to consider the ability of the caregiver. In the dataset, baseline and post-treatment probes display different patterns in terms of the child's attention. Typically, baseline videos have more periods of inattention and little shared attention. Post videos either have more periods of shared attention, or exhibit a clear pattern of transition from inattention to attention, back to inattention, as the parent effectively gained control of the child's motivational activity and releases it after an acceptable response. The effect of the different patterns can be examined by training SVM classifiers on the baseline and post-treatment videos separately. This will allow for an inference to be made regarding how parental familiarity with PRT affects the classification models.

While the first two methods looked at organizing models, the third will look at fine-tuning models to personalize the classification tasks. This will simulate having the clinician review the baseline video to extract labeled data specifically on the dyad that can be incorporated into the model with the expectation that it will improve the classification accuracy in future videos. Intuitively this would be due to two factors. First, recognition of the attributes that distinguish the child and adult in the video could be learned by the model. Second, based on the assumption that the child's preferred activity would appear in multiple videos, the signs of attention particular to that child (or the activity he or she frequently engages in) would be learned by the model. To evaluate this, multiple SVM models were evaluated to look at different ways of incorporating samples from the baseline videos for classification of the post-treatment videos. This effectively reduced the validation size to only the number of samples in the post-treatment videos. As fine-tuning has been a useful tool for computer vision applications, the AlexNet fine-tuning application used in Chapter 3 will again be used as a comparison. This will be trained

with the RGB frames corresponding to the visual pose samples used in the SVM classifiers. In addition to the raw RGB frames, the AlexNet will be trained with cropped images that only feature the hands and faces of the parent and child. This is to prevent the background from affecting the classification performance.

Three experiments were conducted based on fine-tuning the models. First, all of the base video samples were included in the training set, with only the post video used for validation. Second, only the shared and attentive samples from the baseline video are used. This is to help address the data imbalance by including additional samples from the underrepresented classes. Third, the baseline video is classified using a model trained with the videos from the other dyads. Shared and attentive samples that were correctly classified with a high probability were included in the fine-tuning for evaluating the post-treatment video. For SVM implementations, the probability is based on Platt's algorithm as provided by scikit-learn. For the AlexNet implementation, the softmax values are calibrated using temperature scaling (Guo et al., 2017) to create probabilities.

## 6.3 Results

### *6.3.1 Demographic-Based Model Tracks*

The performance results for training separate models based on the perceived vocal ability of the child are presented in Tables 6.1, 6.2, and 6.3. These tables reflect the SVM metrics for using visual pose data, audio-visual data, and only audio data, respectively. Based on these results, no significant difference is apparent when dividing the training set based on the demonstrated vocal abilities of the child. Comparing the attention-based models to the centroid results in Table 5.1 shows that average accuracy and F1 scores remained within a few points of one another. This indicates that data gathered from activities did not differ significantly between the two ability groups. The idea of using the OpenPose data was to

generalize the activities. Although this has not yielded stellar results, the congruency between the methods presented here suggests that there is some generalization occuring.

A more profound effect could have been expected in regard to audio-only results. The separation task results in Table 6.3 are comparable but slightly lower than the values illustrated in Figure 4.7. This could indicate two things. First, the single word group had more class imbalance as these children had fewer utterances, resulting in fewer training and validation samples. The lower representation could have made the child speech more difficult to recognize, lowering the overall validation accuracy of the model. The second indication, which is related to the first, is that the samples of child's speech from the multi-word group were similar enough to the single word group's utterance features to have a positive effect on the classification.

The most profound effect that could be observed from audio classification is in regard to Dyad 6. Dyad 6 represents an outlier in comparison with the children present in the single word group. The child in Dyad 6 was older and articulated words more clearly than the other children in the group; however, he only demonstrated single word responses. This suggests that intelligibility and age could be relevant factors that should be considered when selecting models.

### *6.3.2 PRT Knowledge-Based Model Tracks*

Dividing the data set into baseline and post-treatment video training sets was undertaken to observe if the parent's PRT fidelity has an effect on the attention classification. As stated above, the baseline and post-treatment videos have different attention patterns. Also, presumably, there could be activities that are more common in baseline videos than post-treatment videos, and vice-versa. Similar to separating the dataset by the child's vocal ability, the results of this division were not profound. Examining the visual-feature-only classification model for the baseline videos (Table 6.4) does not exhibit a strong pattern

Table 6.1
The table displays results for using only visual pose features for classifying shared attention, attention (attn), and inattention (inattn). Separate models were used based on the child's demonstrated vocal communication ability, either single word or phoneme attempts, or multi-word speech.

| Model | Validation Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Single Word | Dyad 1 | 0.31 | 0.73 | 0.09 | 0.36 | 0.60 |
| | Dyad 2 | 0.52 | 0.72 | 0.65 | 0.33 | 0.30 |
| | Dyad 3 | 0.36 | 0.72 | 0.26 | 0.39 | 0.39 |
| | Dyad 6 | 0.43 | 0.74 | 0.34 | 0.21 | 0.59 |
| | Average | 0.40 | 0.73 | 0.34 | 0.32 | 0.47 |
| Multi-Word | Dyad 4 | 0.44 | 0.72 | 0.04 | 0.36 | 0.66 |
| | Dyad 5 | 0.44 | 0.73 | 0.43 | 0.20 | 0.69 |
| | Dyad 7 | 0.45 | 0.74 | 0.48 | 0.29 | 0.52 |
| | Average | 0.45 | 0.73 | 0.32 | 0.28 | 0.62 |

Table 6.2
The table displays results for using audio-visual features for classifying shared attention, attention (attn), and inattention (inattn). Separate models were used based on the child's demonstrated vocal communication ability, either single word or phoneme attempts, or multi-word speech.

| Model | Validation Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Single Word | Dyad 1 | 0.36 | 0.98 | 0.11 | 0.35 | 0.62 |
| | Dyad 2 | 0.31 | 0.98 | 0.32 | 0.32 | 0.21 |
| | Dyad 3 | 0.35 | 0.98 | 0.26 | 0.33 | 0.36 |
| | Dyad 6 | 0.48 | 0.99 | 0.40 | 0.26 | 0.64 |
| | Average | 0.38 | 0.98 | 0.28 | 0.32 | 0.45 |
| Multi-Word | Dyad 4 | 0.47 | 0.98 | 0.11 | 0.47 | 0.70 |
| | Dyad 5 | 0.42 | 0.98 | 0.34 | 0.22 | 0.63 |
| | Dyad 7 | 0.34 | 0.98 | 0.32 | 0.30 | 0.43 |
| | Average | 0.41 | 0.98 | 0.26 | 0.33 | 0.59 |

Table 6.3
The table displays results for using audio features for classifying adult, child, and noise samples. Separate models were used based on the child's demonstrated vocal communication ability, either single word or phoneme attempts, or multi-word speech.

| Model | Validation Set | Validation Acc. (%) | Training Acc. (%) | Adult F1 | Child F1 | Noise F1 |
|---|---|---|---|---|---|---|
| Single Word | Dyad 1 | 0.86 | 0.95 | 0.64 | 0.52 | 0.93 |
| | Dyad 2 | 0.77 | 0.94 | 0.64 | 0.27 | 0.94 |
| | Dyad 3 | 0.72 | 0.96 | 0.65 | 0.35 | 0.79 |
| | Dyad 6 | 0.86 | 0.95 | 0.64 | 0.52 | 0.93 |
| | Average | 0.79 | 0.95 | 0.66 | 0.40 | 0.90 |
| Multi-Word | Dyad 4 | 0.70 | 0.95 | 0.75 | 0.41 | 0.76 |
| | Dyad 5 | 0.77 | 0.95 | 0.64 | 0.46 | 0.92 |
| | Dyad 7 | 0.72 | 0.95 | 0.70 | 0.53 | 0.82 |
| | Average | 0.73 | 0.95 | 0.69 | 0.46 | 0.83 |

across the different validation sets. All of the dyads, except 6 and 7, had better accuracy when all of the training data was utilized. Dyad 6 and 7 had moderately better accuracy using only the base training. For Dyad 6, the shared F1 score for the baseline-only training is greater than when all of the probes are used. In this video, the shared attention segments are represented by the child and the parent playing with a toy car track, sitting side by side. A similar scenario is presented in the Dyad 7 video. Conversely, a large portion of the shared attention samples in the post videos are from Dyad 2, and represent the child in the parent's lap watching a movie on a mobile device. These additional samples that have a significantly different structure to the spatial graph could cause confusion of the shared samples when all of the videos are used for training and validation.

The performance of Dyad 2 is the most interesting validation set when examining the visual pose model results (Table 6.6). While the post-only metrics are low (only 21% accuracy) the results when the full training set is used are 61%. As this video depicts the parent and child viewing the movie on the mobile, the sample representation is much more homogenous than other video probes. Since the samples are similar, the classification

models predict a single class for the majority of samples. When the full training set was used, the model benefitted from the other shared attention samples and was able to predict the correct class for most of the samples. Shared attention samples in other post-treatment videos often do not have the same proximity. The activities predominantly include reading a book side-by-side as seen in the Dyad 3 Post video probe or playing a game where the parent and child are sitting across a table from one another. as in the Dyad 7 Post probe.

Examining the audio-visual combined feature sets (Tables 6.5 and 6.7) illustrates that some dyads benefit from the inclusion of audio, while others see a reduction in performance metrics. This is similar to what was discussed in Chapter 5.

### *6.3.3 Fine-Tuning for Personalized Models*

The results for the AlexNet approaches (Table 6.8) illustrates that adding all of the samples from the baseline video improved the validation average accuracy; however, this came at the cost of shared and attention class F1 scores. This indicates that providing the additional labeled samples caused the classification model to select the inattentive class more often than when only the training set was used. This inflated the accuracy and Inattentive F1 scores, but caused the shared and attentive scores to be reduced. The F1 scores were already problematic for shared and attention classes and suggested that the refined AlexNet model overwhelming favors the inattentive class. Adding only the shared and attentive samples from the baseline videos caused a slight drop in accuracy, but an increase in the shared and attentive F1 scores. This illustrates that adding the additional samples aided in the selection of these classes; however, it also caused a greater misclassification of inattentive samples.

Using only the confident shared and attentive samples facilitated a general increase in accuracy compared to using all of the shared and attentive samples; however, the value was still below using all of the baseline video samples. Adding only the confidently

Table 6.4
The table displays results for using only visual pose features from baseline videos on classifying shared attention, attention (attn), and inattention (inattn). For comparison the baseline results for models trained with all baseline and post-treatment videos is presented.

| Validation Set | Training Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Dyad 1 | Base | 0.26 | 0.77 | 0.15 | 0.29 | 0.75 |
| Base | All | 0.27 | 0.73 | 0.10 | 0.34 | 0.32 |
| Dyad 2 | Base | 0.35 | 0.76 | 0.02 | 0.27 | 0.58 |
| Base | All | 0.45 | 0.72 | 0.03 | 0.44 | 0.60 |
| Dyad 3 | Base | 0.34 | 0.77 | 0.30 | 0.37 | 0.33 |
| Base | All | 0.35 | 0.72 | 0.32 | 0.41 | 0.33 |
| Dyad 4 | Base | 0.53 | 0.78 | 0.00 | 0.19 | 0.81 |
| Base | All | 0.58 | 0.72 | 0.00 | 0.21 | 0.74 |
| Dyad 5 | Base | 0.43 | 0.76 | 0.44 | 0.12 | 0.77 |
| Base | All | 0.44 | 0.73 | 0.55 | 0.06 | 0.51 |
| Dyad 6 | Base | 0.47 | 0.79 | 0.45 | 0.16 | 0.51 |
| Base | All | 0.40 | 0.74 | 0.41 | 0.12 | 0.51 |
| Dyad 7 | Base | 0.46 | 0.76 | 0.40 | 0.37 | 0.50 |
| Base | All | 0.39 | 0.74 | 0.39 | 0.32 | 0.45 |
| Average | Base | 0.40 | 0.77 | 0.25 | 0.25 | 0.60 |
| | All | 0.41 | 0.73 | 0.26 | 0.27 | 0.49 |

Table 6.5
The table displays results for using audio and visual pose features from baseline videos on classifying shared attention, attention (attn), and inattention (inattn). For comparison the baseline results for models trained with all baseline and post-treatment videos is presented.

| Validation Set | Training Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Dyad 1 | Base | 0.36 | 0.99 | 0.11 | 0.28 | 0.68 |
| Base | All | 0.39 | 0.98 | 0.11 | 0.35 | 0.52 |
| Dyad 2 | Base | 0.43 | 0.99 | 0.01 | 0.32 | 0.62 |
| Base | All | 0.45 | 0.98 | 0.02 | 0.44 | 0.55 |
| Dyad 3 | Base | 0.33 | 0.99 | 0.25 | 0.29 | 0.33 |
| Base | All | 0.35 | 0.98 | 0.29 | 0.27 | 0.44 |
| Dyad 4 | Base | 0.45 | 0.99 | 0.00 | 0.24 | 0.83 |
| Base | All | 0.51 | 0.98 | 0.00 | 0.28 | 0.44 |
| Dyad 5 | Base | 0.39 | 0.99 | 0.40 | 0.08 | 0.62 |
| Base | All | 0.42 | 0.98 | 0.44 | 0.09 | 0.53 |
| Dyad 6 | Base | 0.43 | 0.99 | 0.50 | 0.17 | 0.52 |
| Base | All | 0.44 | 0.99 | 0.45 | 0.16 | 0.54 |
| Dyad 7 | Base | 0.39 | 0.99 | 0.27 | 0.33 | 0.49 |
| Base | All | 0.40 | 0.98 | 0.30 | 0.32 | 0.51 |
| Average | Base | 0.40 | 0.99 | 0.22 | 0.24 | 0.58 |
| | All | 0.42 | 0.98 | 0.23 | 0.27 | 0.53 |

Table 6.6
The table displays results for using only visual pose features from post treatment videos on classifying shared attention, attention (attn), and inattention (inattn). For comparison the baseline results for models trained with all baseline and post-treatment videos is presented.

| Validation Set | Training Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Dyad 1 | Post | 0.44 | 0.77 | 0.00 | 0.45 | 0.49 |
| Post | All | 0.44 | 0.73 | 0.07 | 0.46 | 0.58 |
| Dyad 2 | Post | 0.21 | 0.74 | 0.34 | 0.00 | 0.00 |
| Post | All | 0.61 | 0.72 | 0.76 | 0.00 | 0.00 |
| Dyad 3 | Post | 0.40 | 0.76 | 0.35 | 0.36 | 0.44 |
| Post | All | 0.35 | 0.72 | 0.19 | 0.36 | 0.45 |
| Dyad 4 | Post | 0.40 | 0.78 | 0.08 | 0.49 | 0.49 |
| Post | All | 0.37 | 0.73 | 0.07 | 0.41 | 0.44 |
| Dyad 5 | Post | 0.51 | 0.78 | 0.25 | 0.35 | 0.63 |
| Post | All | 0.47 | 0.73 | 0.29 | 0.31 | 0.61 |
| Dyad 6 | Post | 0.42 | 0.78 | 0.19 | 0.23 | 0.64 |
| Post | All | 0.49 | 0.74 | 0.16 | 0.23 | 0.66 |
| Dyad 7 | Post | 0.33 | 0.81 | 0.23 | 0.28 | 0.42 |
| Post | All | 0.39 | 0.74 | 0.39 | 0.32 | 0.45 |
| Average | Post | 0.38 | 0.77 | 0.20 | 0.30 | 0.45 |
| | All | 0.45 | 0.73 | 0.28 | 0.30 | 0.46 |

Table 6.7
The table displays results for using audio and visual pose features from post-treatment videos on classifying shared attention, attention (attn), and inattention (inattn). For comparison the baseline results for models trained with all baseline and post-treatment videos is presented.

| Validation Set | Training Set | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Dyad 1 | Post | 0.39 | 0.98 | 0.00 | 0.36 | 0.49 |
| Post | All | 0.42 | 0.98 | 0.02 | 0.40 | 0.54 |
| Dyad 2 | Post | 0.25 | 0.98 | 0.40 | 0.00 | 0.00 |
| Post | All | 0.36 | 0.98 | 0.53 | 0.00 | 0.00 |
| Dyad 3 | Post | 0.36 | 0.99 | 0.21 | 0.36 | 0.43 |
| Post | All | 0.36 | 0.98 | 0.15 | 0.32 | 0.51 |
| Dyad 4 | Post | 0.35 | 0.98 | 0.15 | 0.50 | 0.54 |
| Post | All | 0.42 | 0.98 | 0.12 | 0.50 | 0.42 |
| Dyad 5 | Post | 0.45 | 0.99 | 0.14 | 0.32 | 0.63 |
| Post | All | 0.45 | 0.98 | 0.23 | 0.33 | 0.57 |
| Dyad 6 | Post | 0.45 | 0.99 | 0.11 | 0.28 | 0.67 |
| Post | All | 0.52 | 0.99 | 0.24 | 0.32 | 0.67 |
| Dyad 7 | Post | 0.30 | 0.99 | 0.16 | 0.31 | 0.36 |
| Post | All | 0.30 | 0.98 | 0.30 | 0.27 | 0.32 |
| Average | Post | 0.36 | 0.98 | 0.17 | 0.31 | 0.45 |
| | All | 0.40 | 0.98 | 0.23 | 0.31 | 0.43 |

predicted samples decreased the overall new samples being introduced when refining the model, allowing the inattentive samples to dominate. This is further reflected in the increased inattentive F1 score in comparison to the model trained with all of the baseline shared and attentive samples.

Comparing the images with cropped hands and faces to the full image shows a small increase in the average accuracy and shared F1 score. This could be explained by the differences in the training accuracy. The high training accuracy from the full image could indicate that the model overfit the training set.

Using the visual pose data showed greater overall performance (Table 6.9). The visual pose feature set model had consistent F1 scores through the three training conditions, with a slightly improved validation accuracy when only confident shared and attention samples were used for fine-tuning. Adding the audio features did not aid classification in this scenario in comparison to the visual pose data; however, it illustrated a similar trend. Although not a profound effect, this does indicate that adding additional labeled samples for each dyad positively influences the classification model.

Relying on human experts to label samples that can be incorporated into the system provides a means for introducing error in training the models. The effect of erroneous samples was examined by selecting the same baseline samples used in the confident shared/attention fine-tuning methodology; however, half of the samples were randomly given the incorrect label. The incorrect label that was assigned to the sample was also selected at random. Reviewing the results from both Table 6.8 and 6.9, all of the approaches except the visual-pose-data-only model showed improvement. The F1 scores from the full image AlexNet models suggest that the erroneous data caused the classifiers to assign samples to the inattentive class more frequently. As this class is over-represented, the increased predictions can inflate the accuracy despite more incorrect classification of the other two classes. The audio-visual models' average F1 scores do not

show this trending, with the shared F1 score improved slightly and the attentive and inattentive F1 scores decreased. Explaining this and the larger accuracy increase illustrated with the face-hands-only AlexNet models requires scrutinizing the models' performance on Dyad 3.

As mentioned in previous chapters, the post-treatment video for Dyad 3 is an outlier in the dataset. The samples of this video are overwhelming labelled as shared attention. There is little variation between samples through the video, causing the classifications to be largely concentrated on one of the three classes, particularly for the AlexNet models. For each of the AlexNet model configurations, apart from the face-hand images with random incorrect samples for fine-tuning, the predictions were predominantly in the inattentive category. This caused the validation accuracy to be approximately 1%. For the face-hands models with random mislabeled samples, the validation accuracy rose to 31%, indicating a greater number of correctly classified samples. The explanation for this, and the other increases in accuracy after adding mislabeled samples, could be a reflection of the subjectivity of labeling attention, especially with shared attention.

## 6.4 Discussion

The evaluation presented in the results sections indicates insufficient data to adequately explore dividing the dataset into separate classification models. Intuitively, children of a similar age and communication level are more likely to engage in similar activities and display similar vocal patterns. This was not reflected in the results for the demographic-based models. Other works, such as Rudovic et al. (2018) have demonstrated that it can be an effective way of improving classification. More data is needed to develop more specific models that could be used to personalize the classifications.

Table 6.8
The table displays results for fine-tuning an AlexNet CNN with frames from PRT videos to classifying shared attention, attention (attn), and inattention (inattn). Results are shown for the full image and cropped images showing only the hands and faces of the dyad. Only post-treatment videos were used for the validation set and metrics are averages across the seven validation sets.

| Image Set | Fine-Tuning | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Full Image | Training Set Only | 0.38 | 0.98 | 0.12 | 0.20 | 0.51 |
| | All | 0.41 | 0.98 | 0.10 | 0.17 | 0.53 |
| | Shared/Attn | 0.35 | 0.96 | 0.15 | 0.25 | 0.46 |
| | Confident Shared/Attn | 0.37 | 0.98 | 0.09 | 0.19 | 0.49 |
| | Random Incorrect | 0.39 | 0.98 | 0.08 | 0.20 | 0.52 |
| Face-Hands Image | Training Set Only | 0.40 | 0.89 | 0.13 | 0.17 | 0.50 |
| | All | 0.42 | 0.88 | 0.13 | 0.14 | 0.53 |
| | Shared/Attn | 0.35 | 0.81 | 0.14 | 0.24 | 0.42 |
| | Confident Shared/Attn | 0.39 | 0.87 | 0.06 | 0.18 | 0.51 |
| | Random Incorrect | 0.44 | 0.86 | 0.12 | 0.16 | 0.52 |

Table 6.9
The table displays results for including baseline samples in training to improve classifying shared attention, attention (attn), and inattention (inattn). Two feature sets are presented - visual pose data and audio-visual data. Only post-treatment videos were used for the validation set and metrics are averages across the seven validation sets.

| Feature Set | Fine-Tuning | Validation Acc. (%) | Training Acc. (%) | Shared F1 | Attn. F1 | Inattn. F1 |
|---|---|---|---|---|---|---|
| Visual Pose | All | 0.42 | 0.72 | 0.27 | 0.31 | 0.46 |
| | Shared/Attn | 0.42 | 0.73 | 0.28 | 0.30 | 0.46 |
| | Confident Share/Attn | 0.45 | 0.73 | 0.28 | 0.30 | 0.46 |
| | Random Incorrect | 0.44 | 0.73 | 0.27 | 0.30 | 0.41 |
| Audio-Visual | All | 0.39 | 0.98 | 0.24 | 0.31 | 0.45 |
| | Shared/Attn | 0.40 | 0.98 | 0.28 | 0.33 | 0.46 |
| | Confident Share/Attn | 0.41 | 0.85 | 0.22 | 0.32 | 0.47 |
| | Random Incorrect | 0.42 | 0.98 | 0.26 | 0.29 | 0.39 |

Dividing the dataset by baseline and post-treatment videos was performed to research if the PRT ability of the parent had an effect on the attention classification. This was undertaken because of the clear difference in attention patterns between base and post videos. The performance metrics are mixed, showing some validation sets improved while others did not. The expectation is that this effect would be minimized if following the same methodologies with more data. This can be expected as additional data would introduce more variety that would help the model generalize. Classification in these experiments focused on single frames. This division of the dataset may be more applicable if incorporating sequence-based algorithms that would benefit from learning the different attention patterns.

As with the discussion of the AlexNet performance in Chapter 3, it does not appear that enough data is present to train reliable models based on the RGB data. Increasing the data for each validation set did, however, show an improvement in the F1 scores for the shared and attentive classes when supplying all of the shared and attentive samples from the baseline video. Limiting the number of samples to only those the original classifier was confident with reduced this effect. This could indicate that using a pre-trained AlexNet approach could be feasible with more data. Greater research is needed looking specifically at classifying visual cues of attention based on raw images.

Incorporating the baseline data from the SVM for the visual pose information did marginally improve the classification accuracy. As opposed to other improvements, this did not correspond with an increase in the F1 score of the inattentive class. This suggests that adding the new information helped positively shape the model instead of having predictions cluster to the dominantly represented class. Considering how SVMs are trained, it is not surprising that the smaller number of samples in the confident-only fine-tuning had a more profound effect on the classifier than the corresponding experiments with the CNN-based AlexNet classifiers.

This work was meant to explore the basic concepts of how personalization could be applied using the PRT dataset for a human-in-the-loop feedback system. As such, there are several key limitations that need to be addressed. Classification model parameters were kept static between different feature sets. Adjusting the parameters, such as the number of trainable layers, dropout rate, and learning rate in the AlexNet approach and the C value, game, and kernel in the SVM models, may have produced more optimal results. In particular, SVM parameters were fixed to allow for more accurate prediction probability scores. Not fine-tuning these parameters may have subjected the models to overfit the classifiers. Additionally, the number of different configurations of feature sets and data selection was limited for feasibility. Future work could be undertaken to examine additional organization and selection methods.

## 6.5 Design for Human-in-the-Loop

Implementation strategies for collecting information from clinicians as part of the human-in-the-loop system needs to be streamlined to prevent excessive time costs. Baseline information regarding the child's age and vocal abilities could be easily provided by either the caregiver or clinician during an initial setup. Ideally, in a mature system, the system could automatically assess the child's progress and update the vocal ability designations over time. This would allow the system to adapt and personalize the user experience without additional human intervention. This system could be made more robust by incorporating common assessment metrics rather than a more subjective observation-based designation. Using these formal assessments would provide a better means of clustering individuals and training models utilizing participants with similar qualities. Depending on the protocol the clinician is operating under, this may not require more time investment compared to the current evaluation practices. If the clinician is collecting these metrics as a standard part of their current practice, the only additional

requirement would be to supply the information to the system. Depending on the recording format, this is likely a trivial task.

From a machine learning standpoint, having the clinician involved in system could be interesting for aiding the models to adapt to individuals and become more robust over time. As part of the human-in-the-loop procedure, clinicians are expected to review metrics and media clips of the participating dyad to provide expert advice and feedback. This review process could be utilized to provide additional labeled samples to the system; however, this needs to achieve a balance between the time investment required by the clinician versus the classification performance improvements. The methodology presented above using samples classified with high confidence to be provided as additional training samples was included to simulate a possible implementation for evaluating samples. Instead of requiring clinicians to review a segment in its entirety, individual samples could be selected for approval. This would allow the clinician to quickly accept or reject samples the system identified. As was observed in the results, the effect of mislabeling additional samples was minimal.

In addition to the clinician, the parent could also provide information to the system to aid in personalization of the models. Much of the background assessment and demographic information could be provided by the parents when setting up an initial profile for the system. This could include providing sample video or audio that could be used to calibrate the system or perform preliminary classifications that could be used to personalize the system.

The parent could also review the classification results to determine if an error has occurred or to provide additional labeled samples. This leverages the parent's knowledge of his or her child's behavior in assessing the model's performance. Samples identified as incorrect could be incorporated to fine-tune the model or passed to the clinician for further

evaluation. The parent could also eliminate samples with poor recording quality or disruptive environmental conditions before they are passed to the clinician.

## 6.6 Conclusion

Keeping the clinician as part of the feedback system can be utilized to aid in the personalization of classification algorithms for assessing PRT video probes. This chapter explored different ways that additional information could be added to the system to improve data predictions and present a more optimal experience for the users. The results from fine-tuning an SVM trained on visual pose data suggests this is possible; however, due to limited data, more research is needed to form a firm conclusion on how this can be approached.

Design for incorporating sample label validation from clinicians was presented. This needs to be explored in greater detail and in conjunction with ABA professionals to ensure it is properly implemented. The goal of this aspect of the system should be to minimize the cost of manual assessment of classifier performance. The methodology that was suggested was to choose the samples where the base model was most confident in its prediction, and provide these to the clinician for a binary approve or reject designation. This limits the intervention from the clinician and should not require a significant investment of time.

# Chapter 7

# CLINICIAN USER INTERFACE DESIGN AND EVALUATION

Mapping the current manual data collection processes to automated collection processes requires some consideration. Automatically extracted data collection and analysis can be undertaken at a finer level of temporal granularity. This makes it possible to report metrics for smaller intervals than the one to two minutes used in current practice. Additionally, information that is tedious for human collection, such as number of vocal responses, becomes plausible with automated analysis. How the greater level of detail and additional metrics can be utilized by clinicians remains an open question that this project will address.

Ensuring that an application adequately addresses the needs of its target users requires the developers to consult the users frequently. A prototype user interface (UI) for presenting automated PRT data to clinicians was developed on an iterative life cycle. This consisted of three sprints - the initial design based on literature review, observations of PRT sessions, and conversations with clinicians. The result of this sprint was a wireframe mock of the interface along with examples of extracted data. The data and wireframe were presented as a deliverable to the clinicians at the end of the sprint. The second sprint consisted of building an alpha prototype of the UI. The UI was evaluated using a think-aloud session with the clinicians. The final sprint improved the UI based on the results of the think-aloud session. This beta prototype was presented to the clinicians for review. The primary objectives for creating this UI are to: elicit information regarding how caregiver and child performance metrics should be displayed; determine what new data metrics can be extracted based on the affordances of automated multimodal analysis; and, develop a UI prototype that could potentially reduce the time required for analyzing PRT implementation and providing feedback based on participatory design methodology.

The following sections will discuss the design process in greater detail, and present recommendations on how the UI could be adapted to accommodate caregivers directly.

## 7.1 Related Work

This project incorporates several distinct areas of research. To situate the project into the intended implementation setting, the use of technology in ABA training will be presented. Publications regarding video summarizations and keyframe extraction are relevant to reducing the amount of footage needed for evaluation of the scenario will be discussed. The project design process followed elements of agile software development methodology.

The project incorporates several areas of research. First, research into the application of technology for aiding the ABA training process will be presented. This will provide context related to the intended implementation environment the project is designed for. Next, publications regarding video summarization and keyframe extraction will be discussed. These publications are relevant to the project goal of reducing the time needed to review PRT video probes. Finally, elements of agile methodology will be explored. These elements were incorporated into the design process in order to aid in gaining the clinician's perspective throughout early development.

### *7.1.1 Applications Used By Clinicians for ABA*

Online resources, telemedicine, and video modeling, described in Chapter 2, are technologies utilized in research regarding training individuals in ABA. In practice, two applications were introduced during observations and conversations with clinicians. Naturalistic Observation Diagnostic Assessment (NODA) by Behavior Imaging Solutions is a mobile application designed to assess and diagnose children for autism. The application is advertised for parents to facilitate recording a video on their mobile device to be sent to a panel of clinical experts. The parents interact with the child in a series of

preselected activities. The clinicians then review the video according to the current diagnostic methodologies and return a report to the parents. The application also facilitates requests from the experts for additional videos (Oberleitner et al., 2017; Solutions, 2018b).

This process is based on a human-in-the-loop system relying on asynchronous communication between the parent and the clinicians. Utilizing this process allows the clinicians to review the behavior of the child in a natural context, outside of a laboratory or clinical setting. The application interface was designed to be simple to use. On the client interface designed for the parents, four icons are displayed. These icons relate to the specific scenario that should be videographed. This forces categorization that makes evaluation of the videos simpler for clinicians. For the clinician's interface, the experts can review the videos, create clips, and attach notes to the clips describing the interaction as it relates to the assessment (Nazneen et al., 2017, 2015).

Behavior Connect, also by Behavior Imaging Solutions, is a program and record management system designed for clinical staff. The application allows users to share videos and messages and store client records, with the goals of promoting transparency and facilitating collaboration (Solutions, 2018a).

The research presented in this chapter differs primarily in the incorporation of automated data processing. The aforementioned applications rely solely on human intervention for data labeling and abstraction. The UI presented below examines how the interface needs to be adapted to the clinician's needs based on automated video processing and data visualizations.

### *7.1.2 Video Summarization and Keyframe Extraction*

Manually reviewing videos is a time consuming task, and with the vast amount of video data available, summarization techniques have been an important area of multimedia computing research. Video summarization methodologies focus on detecting differences

between frames, commonly using a histogram representation of the pixel's color value (Sheena and Narayanan, 2015; Thakre et al., 2016). The between-frame comparison results in a distance metric that can be compared against a provided or calculated threshold. If the threshold is surpassed, the frames being compared are from distinct shots or scenes within the video. This can also be accomplished with clustering (Mei et al., 2015) or deep learning (Mahasseni et al., 2017; Zhang et al., 2016) techniques. The end result of this process is the identification of key frames that illustrate an abbreviation of the diversity of the video. An alternative approach focuses on identifying important objects that occur across several frames (Meng et al., 2016).

The problem addressed in this project focuses more on creating a summary of the content in the video, particularly the interaction of the parent and child. These videos are short in duration, around 10 minutes, will likely occur in a single location, and consist of a few distinct actions. Creating automated sports highlights addresses a similar problem. Video highlights in sports have utilized event detection focusing on identifying key actors or action in the frames (Boukadida et al., 2017; Ramanathan et al., 2016). However, unlike the sports examples, the actions depicted in the video cannot be anticipated, since the nature of PRT is dependent on activities selected by the child.

Keyframe detection and video summarization focus at examining the color difference between frames in order to determine when a significant amount of change has occurred to designate a new point of interest. The expectation for PRT videos is that RGB-based approaches for keyframe detection will not adequately capture the important interactions. Significant interactions could occur without a dramatic change in the image representation. Rather than focusing on the visual information for video segmentation, the initial approach utilized for this project was based on audio data. Communication is the fundamental motivation and a large part of clinician evaluation of the caregiver's implementation fidelity. The audio can be used to create segments based on a discrete set

of interactions. The keyframes used to summarize the videos are extracted from the segments created through the audio analysis. In addition to using voice activity for creating video clips, the spatio-temporal data was explored. Graph-based keyframe detection algorithms look at how the spatial patterning of objects in a frame change over time (Demir and Isil Bozma, 2015; Ngo et al., 2005; Vázquez-Martín and Bandera, 2013). This follows a similar methodology to RGB-based keyframe detection; however, by focusing on the change in graph representation of the dyad it was expected to be more interpretive of different interactions.

### *7.1.3 Agile Development*

Agile software development methodologies are based on an iterative delivery of application features. Individual features are designed, implemented, tested, and presented to stakeholders (users and other individuals outside of the development team that are invested in the project) after short development cycles in order to gain end user feedback quickly and adjust future efforts as needed. UI mockups are an important part of this process. Using mockups provides the opportunity to gain feedback from users on the design layout before undertaking development work. Gaining the perspective of the end user on the mockup helps provide insight into the ultimate user experience for the interface (Urbieta et al., 2018).

Ideally, stakeholders should be a fundamental contributor to the design and evaluation of the project. Having a co-design process with the stakeholders will aid in ensuring the project requirements are explicit and being addressed as expected (Kildea et al., 2019). This also helps the project implement a person-centered design, where the application addressed the needs of individuals, as opposed to appealing to a hypothetical ‘average’ user.

Incorporating the stakeholders can be difficult, as it often requires a willingness on the stakeholders' part to invest their time and cooperate over a prolonged period. Often, stakeholders are accommodating in the beginning but become less enthusiastic as time progresses. A recommendation by Urbieta et al. (2018) is to invest more time in the beginning of the process, particularly in examining the user's needs and current solutions to problems the speculative project intends to address.

This project was developed with routine interactions with clinicians to examine desired features, evaluate designs, and test implementations. The application of participatory design toward designing telemedicine or eHealth technologies has not been common (Mitchell et al., 2018). Similar studies in eHealth (Andersen et al., 2017; Gordon et al., 2015) have taken the same approach, as they often rely on a relatively long pilot evaluation period. Gaining cooperation over a long period of time may be difficult given that the application evaluations would likely be in addition to the user's daily work commitments. Following the suggestion of Urbieta et al. (2018), the early mockup development was primarily based on conversations with clinicians, evaluation of current tools being utilized, observations of PRT feedback sessions, and video probe evaluations. This was intended to reduce the time-investment required from the clinicians in the early stages of the project design and development.

## 7.2 Design Process

### *7.2.1 Observations*

Understanding the needs of a system's users is paramount to the design process. Prior to designing a solution, I attended a week-long group training program for teaching caregivers PRT, and observed one-on-one training sessions between behavior analysts and a parent with his or her child. This provided the opportunity to learn about the materials, methodologies, and feedback employed by clinicians when training caregivers. Attending

these sessions also facilitated conversations with parents and clinicians regarding the views on PRT and the challenges of consistent long-term usage. The primary observations from these sessions relevant to this project are: performance feedback from clinicians is situated in the context of an activity; evaluating video probes for fidelity was time-consuming; feedback on progression relied on manually gathered data; and post-treatment feedback and support was minimal. These observations formed the base assumptions utilized for the initial UI design.

During the observed one-on-one sessions, the caregiver practiced implementing PRT with her child in the presence of the clinician. This afforded the clinician the opportunity to provide suggestions and feedback in real time. This benefits the caregiver by helping her situate the feedback into the context of his or her immediate behavior. It allows the caregiver to immediately act on the suggestion. Commonly, the feedback that was provided started with example. Either the clinician modeled a behavior with the intention that the caregiver emulated that behavior, the clinician praised a specific interaction that occurred, or the clinician identified a particular instance where the caregiver acted incorrectly or missed an important opportunity.

Only the adults were present during group sessions. Clinician evaluation during these sessions was conducted using 10-minute videos of the adult and child interacting. The videos were reviewed as a group, with the clinician pausing to provide feedback. As with in-person training sessions, this allowed the clinician to isolate specific instances in the video where the feedback was applicable.

Outside of providing face-to-face feedback, scoring video probes is a manual task, requiring a clinician to review the video, evaluate the adult's behavior in regards to implementation criteria, and identify frequency of child vocalizations. This involves watching the video probe multiple times to ensure proper assessment. For adult implementation, the scores were assessed in minute increments on a binary system

reflecting if the adult adequately met the criteria during the interval. The criterion is met if the adult correctly demonstrates the behavior twice during the interval. No data is recorded on the specific action that was taken or where in the interval the actions took place. The scores for each category are tallied and averaged to create a fidelity score percentage. Achieving a score of about 80% is considered adequate implementation.

Primarily, utterance frequency is used for child vocal assessment. This is based on a presence or absence value determined by whether or not the child vocalized during a 15 second time period. This can be recorded on whether the utterance was spontaneous or a response to an instruction from the caregiver, however, this practice was not observed. The only information collected on the child's vocal attempts were binary indications of an adequate vocal attempt being made.

Upon completing the one-on-one training, the caregiver is presented with a report detailing the treatment and the evaluation metrics from the video probes. This presents a comparison between an assessment of a baseline video probe recorded prior to receiving training, and a post-training probe record on the final day of the course.

After the course is concluded, options for continual support are limited. As observed, the process of training and evaluating caregivers is intensive, and centers lack the resource availability for support after training. This is problematic. In interviews, clinicians informed me that one of the primary challenges that led caregivers to abandon PRT is an inability to adapt the procedure to new activities and learning objectives. This was also related to me during the group course by a participant that was attending the class after previously undertaking the one-on-one training.

### *7.2.2 Automated Data Processing*

Both the parent's implementation fidelity and the child's vocal ability improvement is determined by evaluating baseline video probes recorded before treatment and

post-treatment video probes. Data automatically extracted from each video could be used for the same purpose. Automated analysis focused on multimodal data, utilizing the video and audio recorded in the video probes. The probes present a challenge for automated processing and classification. The videos are often recorded using a handheld camera or mobile device, causing visual instability. Additionally, due to the activities that are commonly depicted, the individuals in the frames are often partially or fully occluded. Audio processing is also challenging, particularly for identifying child vocalizations. The audio recording quality is often dependent on the camera operator's proximity to the dyad, and external noises. Additionally, all child utterances need to be identified. Depending on the vocal ability of the child, the utterance may represent only attempts at speech that may just include individual phonemes.

For the initial research, automated data processing focused on extracting information about the attention state of the child and identifying the caregiver and child vocalizations. Extracting the attention state of the child follows the methodology described in Heath et al. (2018). The video was processed using OpenPose to detect body and facial landmark points from the individuals in each frame. This data, along with an estimation of the visual focus of the individual, was used to construct a spatio-temporal graph of the dyad in the frame. The data was used to train a support vector machine (SVM) classification model based on three class labels - attentive, inattentive, and shared attention. The labels were based on 30-frame segments, representing approximately one second of the video. The conclusion reached in Heath et al. (2018) stated that this method only produced 44% accuracy on individual frames; however, an accuracy of 56% is achieved when aggregating samples to assign a label to 30 frame segments.

The audio assessment was based on research presented in Heath et al. (2019c) using the same dataset as Heath et al. (2018). The probes' audio data was processed to extract common features using PyAudioAnalysis, and was classified as being silence, noise, adult

speech, or child vocalizations. Two SVM models, one trained to separate speech from noise, and a second for differentiating child and adult vocalizations, were used for the classification tasks. Using this methodology, an overall classification accuracy of 79% was achieved.

### *7.2.3 Video Clip Creation*

Segmenting the videos into semantically meaningful clips was identified as an important feature for utilizing automated processing. This would reduce the amount of time the clinician needed to invest in viewing the video and extracting example clips for situating pointed feedback. These clips were created by first analyzing the audio data to identify when adult and child vocalizations occur. This classification information was then used to create segments containing adult-only speech, child-only speech, and both adult and child speech. Based on the temporal relationship between the adult and child speech, the child speech could be labeled as either spontaneous or a response. Each of these segments were then classified for the attention state by analyzing the video frame data of the segment and using the most represented label.

### *7.2.4 Initial Assumptions and Project Goals*

The objective of the project is to explore how data metrics automatically collected from video probes could be used to reduce the amount of time clinicians need to invest to evaluate caregiver PRT implementation, and how an interface can be designed to afford new opportunities for clinicians to provide feedback. This involves not only how the interface can be designed to promote ease of use, but also how PRT evaluation procedures can be expanded by new data collection techniques.

Initial assumptions on the design were developed to alleviate the need to view the entire video. Creating the video segments would allow clinicians to view an annotated

storyboard of the video and select the clips they felt were important to view and remark on. Additional information would be provided in accompanying graphs. These graphs were expected to provide a summary of the video, or sections of the video, that could be utilized to gain an understanding of the interactions prior to reviewing the clips. Additional graphs would be made to display comparison data between video probes from the same dyad. This was intended to show the progression of the caregiver's PRT implementation fidelity and the child's vocal communication skills.

The data for the graphs was related to the evaluation criteria that the clinician manually collected; however, at this stage it was not intended to supplant the manual metrics. The data that was initially thought to be important to track was the overall percentage of each attentive state in the video, the percentage of audio classification throughout the video, the number of adult speeches that occurred while the child was inattentive or attentive, the number of child spontaneous vocalizations and responses, and the average length of the child's utterances.

### *7.2.5 Sprint 1: Wireframe*

The wireframe created for the first sprint consisted of a mockup of the primary interface page and graphs for visualizing the data extracted from the video. The primary interface page (Figure 7.1A) was designed to function as a storyboard depicting an abbreviation of the video along with controls for launching media players and filtering the gridview based on the video clip labels. Each row of the gridview consists of the time interval, audio label, attention state label, four screen shots from the clip, audio play button, video play button, and a comments section. The screen shots were selected from the clip at evenly spaced intervals. This is intended to show the interaction, without the necessity of viewing the full video. The comments field was intended for clinicians to provide feedback to the caregiver based on that specific clip and allow the feedback to be situated within its context.


Figure 7.1: Depicts the storyboard page throughout the project. (A) is the wireframe mockup, (B) is a screenshot of the alpha build, and (C) is a screenshot of the beta build. Video images were retrieved from (Virgir05, 2015)

Pie charts, bar graphs, and box plots were chosen for displaying data metrics extracted from the videos. The percentage of the video pertaining to specific attention states and audio labels was illustrated in pie charts (Figure 7.2). These charts were created to provide a broad overview of the video. In accordance with current procedures of evaluating each minute of the video, bar graphs (Figure 7.3) were utilized to show important related vocal utterance scenarios for the child and parent. In particular, these show when the parent is vocalizing based on the attention label and if the child vocalization is spontaneous or a response. A vocalization was determined to be a response if the child vocalized within three seconds after the adult spoke. More information on the child utterances were presented in box graphs. Bar graphs were created for each minute along with a total graph representing the cumulative data from the video.

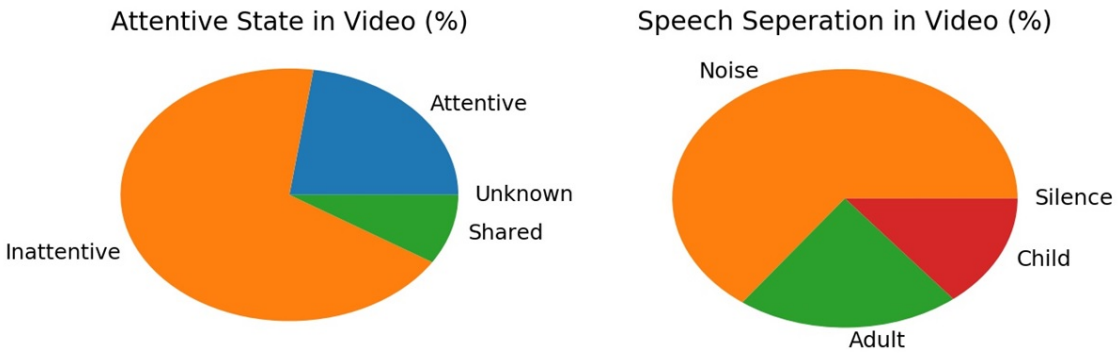


Figure 7.2: Example pie charts showing child attentive state and speech separation from video analysis.

Mean length of utterance is a metric that is commonly seen in PRT research as a measure of the child's vocal usage. This was represented as a box and whisker plot (Figure 7.4) showing the distribution of the child's vocalizations. At this stage, these vocalizations were based on aggregate 250 ms segments as classified by Heath et al. (2019c). Frequency of the child utterances was displayed in a grid-based plot showing the presence or absence of at least one vocalization in a 15 second interval (Figure 7.5).

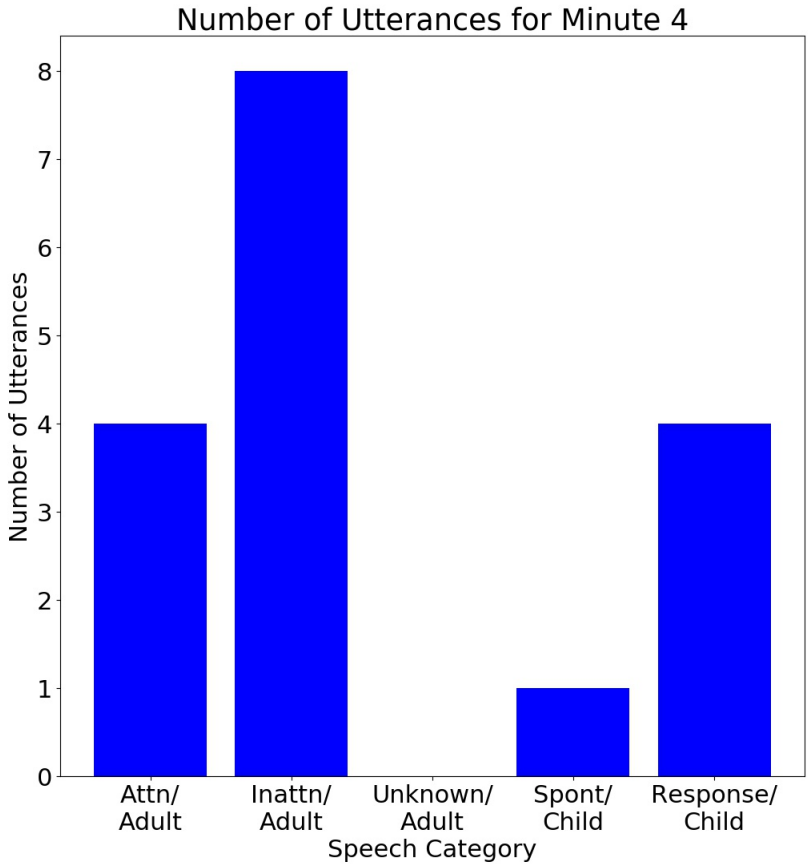


Figure 7.3: Example bar graph showing the number of vocalizations in a one-minute segment. Shows adult speech based on child attention state, as well as child responses and spontaneous utterances.

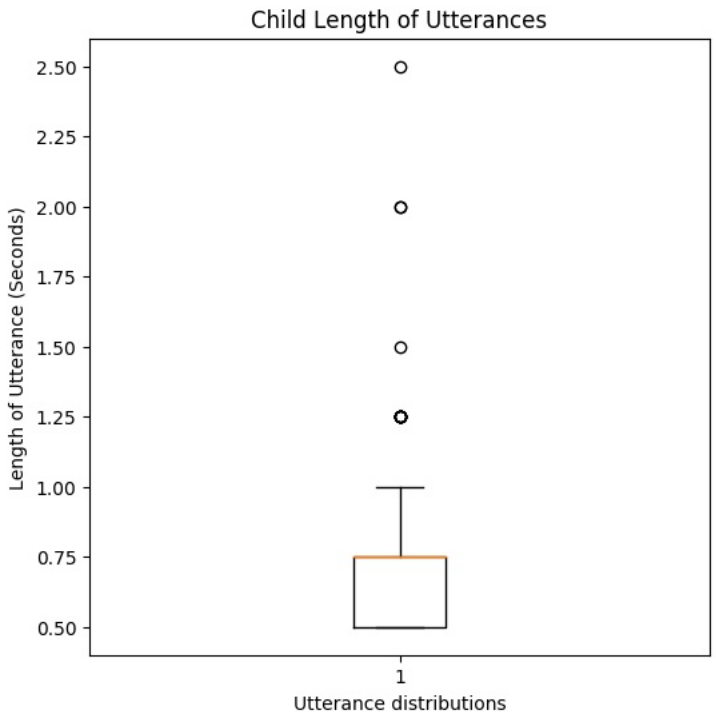


Figure 7.4: Box plot showing the length distribution of the child's vocalizations. In this plot, the mean is the top boundary of the box.

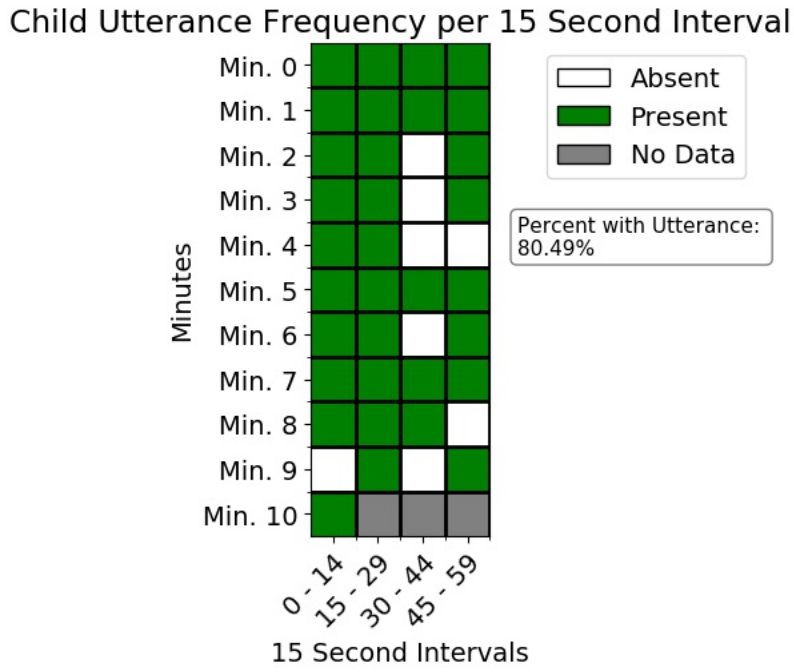


Figure 7.5: Grid showing the presence or absence of a child vocalization for each 15 second increment of the video.

The storyboard, graphs, and media clips were presented to a group of four behavior analysts and the design was discussed as a group. Design critiques focused primarily on the storyboard layout. The use of four screenshots for each segment was seen as superfluous. The segments typically encapsulate a few seconds of the video, leading to little new information being provided in each shot. It was stated that a single shot would be sufficient, as the most important information in the image was the toy or activity the dyad were engaged with.

Buttons on the storyboard page were also reorganized. The audio-only button was not seen to be beneficial. With the reduction of the screenshots to a single image, it was requested that the image be used to trigger playing the video clip. Additionally, a button for playing the video in its entirety was desired.

The behavior analysts also wanted to be able to easily view the interaction in context. To facilitate this, they wanted the ability to continue watching the video after the clip's ending as well as view preceding seconds of the video without having to navigate to a different segment. The ability to save these adjusted clips was also requested.

#### *7.2.6 Sprint 2: Alpha Prototype*

After receiving the feedback from the sprint 1 deliverable session, the UI was developed using a .NET WPF framework. The storyboard page (Figure 7.1B) was developed using a dynamic grid that loaded screenshots and video segment information created in a separate video extraction process. The framework supported column sorting, resulting in removal of filter buttons. This page supports multiple videos of the same dyad by using separate tabs for each video, allowing the clinician to easily move between them. For each segment, an export button was added to the row to allow the clinician to save the video clip. As requested, a button for playing the full video was added to the top of the screen.

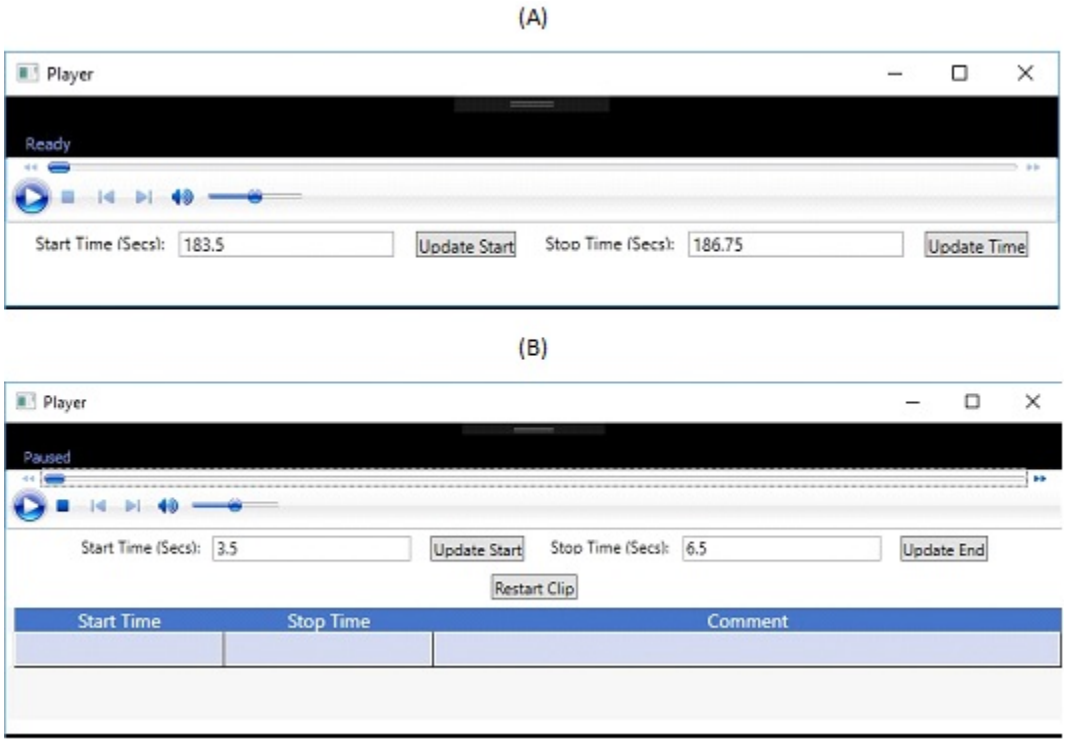


Figure 7.6: Screenshots of the video viewer window. (A) represents the alpha build, while (B) is from the beta build. The pane for viewing the video has been reduced.

An additional page was created for displaying the video segments (Figure 7.6A). This page is launched when a user clicks on the screenshot image. When launched, the page loads the full video and cues the playback pointer to the segment start location and begins playing the video. Once the segment end point is reached, the video is automatically paused, but can be continued by pressing the play button. Pressing the 'Update Start' or 'Update End' buttons will save the current playback time as either the start or end time respectively, allowing the clinician to re-cut the clip.

The deliverable session of the alpha prototype was conducted using a think-aloud methodology (Lewis and Rieman, 1993) with the same group of clinicians as the previous session. During the think-aloud, one of the behavior analysts walked through the use of the UI while verbally describing her actions. This provided perspective on the flow of events that could be expected in a typical use-case scenario. She began by viewing the

entire video to perform the typical fidelity scoring process, making a note of exemplary interactions. Next, she viewed the graphs, particularly noting the ratio of inattentive versus attentive states. Finally, she examined the video clips, stating she would first look at attentive and shared attention samples, then view inattentive samples. While going through the think-aloud it became apparent that clinician's notes that were not intended to be passed to the parent were an important part of the process and should be accommodated in future designs.

Using the video segment viewing page also elicited areas to improve and expand the UI. During the think-aloud, the clinician viewed the clip, then extended the segment past the end pointer. She then created a new end point and wanted to return to the start of the clip. This involved viewing the start time displayed in the text box and manually moving the playback pointer to the appropriate spot. This could be improved upon by providing a button to reset the cursor to the beginning.

An additional feature that was requested was the ability to add comments directly to the video in addition to external feedback that pertained to the entire clip. These notes would be intended to further capitalize on situated feedback by displaying the comment for the duration of the video it is applicable to. This will help the feedback recipient to contextualize the information within the specific action.

#### *7.2.7 Sprint 3: Beta Prototype*

For the third sprint, the storyboard page (Figure 7.1C) only received minor revisions. In the video segment grid, the export button was removed and a new, editable text field was added for clinician notes. A single export button was placed at the top of the page for creating and saving video clips. Additionally, a separate button was created for displaying single video graphs and graphs comparing multiple videos.

The major features for the beta were centered around saving data and exporting video clips. For this version of the project, comments and notes were saved to a file, while video clips were created and saved in a specified directory. Only segments that have comments from the clinician are selected for export.

The video clip viewing page (Figure 7.6) was updated to include a button to restart the video from the clip start position and a grid was added for creating in-video comments. This grid allows the clinician to input the start and stop time that a comment will appear on the video. The text overlay consists of a dialog box with high transparency to prevent obscuring the video. This can support multiple comments appearing at different times during the video clip.

The beta prototype was evaluated by providing the same group of clinicians as the previous sprint with a downloadable copy of the application. The download package included the program files, data, example results, and a text document containing installation and usage instructions. The clinicians were given 12 days to review the application, and were asked seven questions regarding the usability of the program, the practicality of the information being provided, and desired features not included in the prototype. Due to the small number of evaluating users, responses was not anonymized.

Overall, the clinicians stated that the program was easy to use. The clinicians felt that navigating between videos, launching segments, and viewing graphs was intuitive. The installation process for the prototype required mapping directories to load the data correctly, resulting in unnecessary and confusing installation steps. This would be improved in future iterations by using installers and distributed data storage systems.

The clinicians were excited by the data that was being provided and felt that it would reduce the amount of time necessary for reviewing the videos. They felt that this would aid in the feedback they could present caregivers. Additionally, the ability to add comments directly to the videos was a praised feature. In regard to data, the clinicians

stated that the metrics being extracted reflected the majority of what would be useful, and did not have suggestions for additional information that should be collected. It was stated that providing the child's utterance frequency data (Figure 7.5) would reduce the time needed to review a video by 10 to 20 minutes.

No new features were suggested for future implementations, however, there were requested improvements. One of the questions asked referred to the minimum segment size for video clips the clinician felt would be useful. Each clinician stated that 10 to 15 seconds is the smallest increment they would like to see. In its current state, the smallest clip size is 0.75 seconds. Accommodating this will require examining sequence classification and result aggregation in future research. Apart from segment size, quality of life improvements, such as providing numerical information on graphs and more descriptive graph titles, were requested.

## 7.3 Discussion

The goal of this project was to gain an understanding of how automatically collected data could be utilized in the feedback and evaluation of caregiver-implemented PRT sessions. This required looking both at ways to facilitate the current data collection and evaluation process, and how it can be expanded. The current process collects minimal data based on segments of time due to the cost of human evaluation. Automated processing allows for fine granularity of data to be collected that would aid in providing overviews of the videos along with facilitating comparisons between videos. The intended effect would be to provide long-term data tracking that could be used to identify progression and indicate when skill acquisition plateaus are encountered. This would aid the clinician in providing encouraging feedback, and indicate when new skills or approaches need to be introduced. Based on the feedback sessions, the data graphs created for the prototype would be useful in accomplishing this goal.

The progression of the user interface illustrated that several of the assumptions that were involved in the initial design did not come to fruition. The goal was to eliminate the need to review the entire video, however, the first action that was taken during the think-aloud was to watch the entire video and perform the traditional evaluation. This is understandable to an extent at this stage of the project, as the metrics being automatically collected do not encompass the entirety of the criteria being used for evaluation by clinicians. The segmentation layout and automated collection of child utterance statistics did eliminate the need for multiple viewings of the video.

An additional assumption that was false regarded the still frames used for the storyboard. It was presumed that the clinicians would benefit from multiple images in order to gain an understanding of the actions undertaken by the dyad in the segment. The clinicians stated that this was unnecessary as the actions could largely be deduced from the visible objects in the frame.

Also in regards to the segments, the initial assumption was that isolating speech events would be the preferred method for creating each clip. While this is partially correct, the clinicians each stated that this information alone was inadequate. The group consensus was that clips should be at least 10 to 15 seconds long, encapsulating the important events. The context leading up to the vocal event, and the consequences after, were valuable for feedback, as well as creating more concrete examples of behaviors.

The two best received features were the frequency of utterance graph and the ability to create customized video clips. The frequency of utterance graph is a full automation of a currently undertaken task. The clinicians saw a direct benefit to this information since it would no longer be needed to be collected by hand. Similarly, the clip controls were seen as valuable for the creation of demonstration and training tools. During the think-aloud session, the first comments on how this could be useful were related to creating examples

for training and presentations, rather than the effect it would have on reducing the time required for conducting evaluations as intended.

The most important outcome of this project is that it forms the basis for continuing to integrate automated data collection into PRT evaluation. The data being captured in the graphs was based on a conceptualization of what would be important to provide more specific feedback at a reduced human cost. Apart from the utterance frequency metrics, this data is not currently being collected for use in caregiver training. The project provided an opportunity to evaluate how useful this data could be, and the best way it could be presented to clinicians to increase the feedback that can be provided to caregivers.

## 7.4 Re-examining Keyframe Detection and Clip Automation

The audio signals were used for creating video clips; however, the clinicians felt the granularity of the clips was too small to be of specific use. Instead they want 10 to 15 second clips that encapsulate the interaction. Naively, this could be accomplished by localizing the audio signal, then expanding the clip to include preceding and succeeding videos. Depending on the actions that are depicted, this may not be the ideal division. Examining keyframe detection techniques could provide a method of ensuring that the important interactions are depicted in the video clip. Detecting the keyframes would provide starting and stopping points for video clips. Before incorporating the audio signal, it was important to evaluate how current keyframe approaches perform on the PRT dataset. There are two related goals that keyframe algorithms address. First, keyframe detection has a low threshold of change and is intended to abbreviate the video by removing frames with a high similarity. This was explored on the PRT dataset using Key Frame Detector[1]. The second approach, scene or shot detection, has a higher threshold for change and is more concerned with detecting larger background changes that indicate a scene change in

[1] https://github.com/joelibaceta/video-keyframe-detector

the video. Scene detection was applied using Shotdetect[1]. Both Key Frame Detector and Shotdetect compare data frames using the histogram of pixel color values.

Neither keyframe detection nor scene detection are ideal for identifying the important interactions in the videos. The PRT sessions are shot in a single environment and may not contain significant changes in the pixel values to suggest a different scene. For the most part, this is reflected in the results from using Shotdetect on the PRT videos, where the majority of the videos have under five shot change frames detected. The high results from the Dyad 2 Base video are understandable considering the child in the video is highly ambulatory and there is substantial movement of both the individuals and the camera throughout the video. The high number of detected scenes from the Dyad 2 Post and Dyad 4 Post videos are more difficult to explain. In the Dyad 2 Post video, the parent and child are seated and watching a video on a mobile device. There is some instability with the camera and the child changes seating positions; however, there are no other major frame changes that would indicate 19 different scenes. The Dyad 4 Post video is a similar scenario with the parent and child playing with toy cars on the floor. The entire scene has a singular perspective of the interaction. Making the threshold value for detecting scenes more sensitive did not change the number of keyframes that were detected.

Key Frame Detector responded to threshold adjustments. Keeping the value too low, at 0.3, resulted in thousands of keyframes for each video, while a high value, 0.8, produced more manageable results. These capture the macro movements of the individuals, but it is difficult to calibrate to account for more subtle movements. Making the threshold too sensitive causes the algorithm to react to camera instability.

A graph-based approach to keyframe detection is likely to yield more favorable results. Part of the process involves extracting the body pose points for the individuals in the frame. These values can be used to detect when substantial body movements occur

[1] https://github.com/yu239/shotdetect

that indicate a change in the interaction. This change can be used as the basis for detecting keyframes. Using this approach will aid in addressing camera instability and occlusion that could be problematic for interaction-based keyframe detection.

Part of the data extraction process for the body pose information incorporated methods for normalizing the data. The normalization process was intended to account for different frame sizes; however, this also provides an anchor point that addresses instability. The individual on the left of the frame's neck coordinates are used as a common origin for the body points from each individual. All of the point values will be based on their relation to this point, making the camera perspective less influential than would be expected in a color histogram approach.

Occlusion was also addressed in the pre-processing techniques described in Chapter 3. This approach attempts to approximate a missing point's location based on past and future values. Apart from this, if a body point is occluded from the graph it should not be utilized when comparing two frames. This was accomplished by removing the missing point and the corresponding point in the comparison frame from the calculation of the similarity score.

Reviewing the keyframes counts in Table 7.1 provides insight into how the different approaches compare with one another. Primarily, what needs to be determined is how well the algorithms are able to detect important interactions between the parent and child in the video, and its robusticity to camera instability and occlusion. For these we want to look at the most dissimilar results in the table. Intuitively, the dissimilarity identifies videos where the approaches diverged in the information that was useful for determining keyframes. The Dyad 2 Post, Dyad 4 Post and Dyad 7 Post videos provide scenarios that can be used for diagnostics.

The Dyad 2 Post video depicts the child and parent watching a video on a cell phone. As described previously, there is no substantial movement or shift in perspective during

Table 7.1
The number of keyframes detected for Shotdetect and Key Frame Detector are compared to an ST graph-based approach. Results are reported for multiple threshold values.

| | Shotdetect | Key Frame Detector | | | ST Graph | | |
|---|---|---|---|---|---|---|---|
| | 1.0 | 0.3 | 0.6 | 0.8 | 300 | 400 | 500 |
| Dyad 1 Base | 3 | 3366 | 70 | 17 | 60 | 26 | 10 |
| Dyad 1 Post | 3 | 4521 | 390 | 47 | 149 | 88 | 66 |
| Dyad 2 Base | 3 | 3042 | 301 | 83 | 86 | 50 | 34 |
| Dyad 2 Post | 27 | 5714 | 252 | 26 | 18 | 5 | 4 |
| Dyad 3 Base | 19 | 4044 | 182 | 34 | 180 | 132 | 87 |
| Dyad 3 Post | 1 | 2792 | 60 | 8 | 84 | 46 | 27 |
| Dyad 4 Base | 5 | 1898 | 87 | 12 | 45 | 31 | 16 |
| Dyad 4 Post | 35 | 2676 | 127 | 29 | 139 | 79 | 53 |
| Dyad 5 Base | 1 | 5682 | 188 | 32 | 132 | 83 | 49 |
| Dyad 5 Post | 1 | 6006 | 746 | 25 | 249 | 211 | 176 |
| Dyad 6 Base | 2 | 5903 | 555 | 14 | 114 | 73 | 61 |
| Dyad 6 Post | 2 | 5610 | 306 | 41 | 329 | 275 | 218 |
| Dyad 7 Base | 1 | 5600 | 284 | 28 | 165 | 126 | 95 |
| Dyad 7 Post | 2 | 5931 | 425 | 7 | 104 | 78 | 53 |

the recording. Interestingly, both the Shotdetect and Key Frame Detector have relatively high keyframe counts in comparison to their performance on other videos. Conversely, the ST graph approach has a much lower keyframe count compared to other videos. Reviewing the video, there is steady camera movement likely reflecting the breathing of the operator, and occasionally more erratic movement when the operator adjusts her position. These movements are not substantial; however, because the comparisons are relative to the video and there is little movement from the individuals and no changes in the background, the camera movements are triggering the keyframe detectors. Confirming this, the images from the keyframes map to periods of camera movement. As hypothesized, the ST graph was less influenced by the camera movement, and the keyframes are limited to the movement of the dyad. A similar effect is likely why Shotdetect identified a high number of keyframes in Dyad 4 Post. In the majority of the video, the individuals are sitting on the floor playing with cars. Unlike in Dyad 2 Post, there is movement from both the parent and the child as they drive the toys along the ground with their arms and maneuver into different positions. These actions seemed to be adequately captured by the ST Graph approach and, depending on the threshold, the Key Frame Detector. However, the Key Frame Detector appears susceptible to camera instability, particularly in the early part of the video.

The Dyad 7 Post video begins by recording the parent and child selecting a game from a counter, then sitting at a table to play the game. What is interesting in the keyframe metrics is the difference in the Key Frame Detector performance between threshold values of 0.6 and 0.8. Reviewing the frames from the 0.8 threshold reflects the process of selecting the game, moving to the table, and sitting down. No additional frames are reported after sitting, despite periodic hand and arm movements when each individual takes his or her turn. These movements were then captured in minute detail at a threshold

of 0.6. The ST graph was more stable between threshold changes and was able to capture more of the key movements while the dyad are playing the game.

Viewing the results for Dyad 3 Post and Dyad 7 Base provide insight into how the detectors perform under occlusion conditions. Dyad 3 Post is an example of periods of partial occlusion, particularly in the middle of the video when the parent is showing an illustrated book to the child. The parent is often partially occluded, including frames where her face is blocked by the book. The algorithms appear to be fairly resilient to this behavior. Dyad 7 Base has sustained periods where the adult is not in the frame. During these periods, the ST graph was more sensitive to creating unnecessary keyframes.

These results suggest that more dependency on the individuals' movements reflected in the ST graph is a viable method for addressing camera instability; however, more research is needed, particularly in addressing full occlusion. This research would include evaluating different similarity metrics such as cosine distance or Mahalanobis distance if using a clustering approach. The Key Frame Detector approach may also be viable if provided an optimal threshold. These approaches also need to be validated using additional datasets depicting dyadic interactions.

## 7.5 Extending the Interface for Parent Usage

The application evaluated in this article was designed to accommodate clinical professionals and aid in the support they are able to provide caregivers learning to implement PRT. The question of how relevant this system would be if provided directly to the caregiver is worth addressing briefly. Caregivers choose to learn PRT in order to gain a structured methodology for helping their child develop social and communication skills. Acquisition of knowledge on intervention methods is left solely to the caregiver, requiring him or her to be self-motivated to improve implementation of the techniques. Because of

this, designing technology to directly aid parents learning PRT needs to follow principles of self-regulatory learning.

Self-regulatory learning has four primary phases: defining a task, setting goals, working on achieving the goals, and adapting the process to a new task (Winne, 2011). Utilization of the storyboard features presented in this article could aid the learner in evaluating progress on meeting his or her goals. By automatically creating segments of interest, the caregiver could review specific areas of the video that could highlight correct implementations of PRT, while also showing missed opportunities or ineffective behaviors. Reviewing this video immediately after the PRT sessions would help the caregiver situate the feedback in the context of the activity (Robinson, 2011; Suhrheinrich and Chan, 2017). This provides a framework for self-monitoring and self-evaluation of the caregiver's interaction with his or her child (Kitsantas and Kavussanu, 2011).

Building self-efficacy is an important part of the learning process and is a key component of successful self-regulatory learning (Kuhl, 2000). This situation is an interesting scenario from a learning and motivational perspective. The caregiver is primarily interested in the knowledge acquisition and improved skill performance of the child, not his or her own; however, the child's improvement will be influenced by the caregiver's implementation fidelity. Fostering self-efficacy in the caregiver, and improving his or her confidence in the effectiveness of PRT, is dependent on both the child and caregiver performance metrics and how they progress over time.

Although self-regulatory learning is directed at the individual, facilitating communication between the learner and experts is an important part of the process. Learners who have access to other individuals to ask questions are more likely to be successful (Pintrich, 2000). In the case of PRT, maintaining a connection between the caregiver and the clinician will help the caregiver gain confidence in the implementation

strategies, as well as provide a resource to aid in overcoming plateaus in the learning process for the caregiver and the child.

## 7.6 Future Development

The UI presented in this article reflects a prototypical design evaluating core components that would facilitate evaluation and feedback tasks for behavior analysts. Further work could be undertaken to make this prototype more robust and personalizable.

This UI is intended to be part of a larger system. A complete system would include a persistent back-end for data storage, a distributed architecture for handling the video processing tasks, and a second front-end application for caregivers to upload videos, receive clinician comments, and view automatically extracted performance metrics. Designing the caregiver UI would benefit from a similar approach to the clinician UI, incorporating both clinicians and caregivers to evaluate the UI at incremental stages in the development. The expectation is that caregivers would benefit from features that help provide motivation and support self-regulatory practices.

More effort is needed to improve the automated video clip creation and labeling. Based on clinician feedback, the video clips need to be aggregated to a minimum length of 10 to 15 seconds. Additional research will focus on how to evaluate longer sequences of video data, as well as examine multimodal solutions to identifying periods of attention.

In addition to the current data being extracted, there is the opportunity to provide information regarding the affective state based on video, audio, and lexical data (Bertero et al., 2016; D'mello and Graesser, 2010; Le et al., 2017; Parthasarathy and Busso, 2017; Rudovic et al., 2018). Data on the emotional state of the individuals could provide useful information to the clinician, as well as being important for caregiver self-reflection on depicted activities. This could also provide insight into activities that are particularly motivating for the child.

The design evaluation for this project was limited to four behavior analysts. This provided adequate insight to develop a prototype; however, the application would benefit from recruiting a broader number of users for future evaluations.

## 7.7 Conclusion

Incorporating behavior analysts into the design and development process afforded the opportunity to gain a greater perspective on not only the interface needs of the project, but the practicality of the information being collected and presented. Following the collaborative design structure in combination with aspects of agile methodology allowed the researchers to gain perspective and pivot the project to meet the clinician's needs as they were identified. The foundation of the project, the initial wireframe, was based solely on observation and research. Although fundamentally useful, it was immediately apparent to the clinicians that it needed to be simplified. Utilizing a think-aloud for the evaluation of the alpha prototype created the opportunity for both researchers and clinicians to identify missing key functions and desirable features. The final evaluation of the beta prototype gave the clinicians the opportunity to experience how the application could be utilized, and provided important feedback for continuing the project in the future.

Additionally, the clinician feedback drove the need to discover methodologies for capturing important interactions in a video clip. Using color histogram-based approaches to keyframe detection were subject to camera instability. To address this, a keyframe detection approach using the ST graph of the visual pose data was introduced. Preliminary results suggest this is a more robust method for determining keyframes that could be used to create video clips.

# Chapter 8

# CONCLUSIONS, UNANSWERED QUESTIONS, AND RECOMMENDATIONS FOR FUTURE WORK

This dissertation has evaluated the initial steps for designing and creating an automated feedback system for supporting parents learning to implement PRT with their children. This domain exposes novel and complex problems for machine learning and artificial intelligence research. The forefront of the research discussed in previous chapters examined detection of child attention. Detecting human behavior is a distinctly more complicated problem than detecting specific activities. Human behavior manifests differently in different scenarios and activities, and often varies between individuals. The research presented sought a means of using machine learning techniques to generalize the detection of attention to support the variation that can be expected in PRT video probes. Accomplishing this task has important applications in a wide variety of domains ranging from education to public safety.

PRT video probes offer a novel scenario for exploring speech-related technologies. The difficulty of common VAD and ASR techniques to model the child-directed speech patterns utilized by the adults in the PRT videos reflects models that are rigidly trained under assumptions of acceptable degrees of articulation. This illustrates challenges that need to be addressed in order to accommodate people with atypical or varied speech patterns. The PRT videos also present a situation where it is not only important to detect and understand verbal speech, but where nonspeech vocalizations are equally as important. This is reflected in the need for the speech system to correctly recognize all of the child's utterances during the interaction.

In addition to these challenges, the recording environment presents obstacles that are relevant to the application of machine learning in the real world. Recording environment

and quality of data capture cannot be guaranteed in the PRT video probes. Correctly assessing the videos requires using methodologies and algorithms that are robust to missing and noisy data.

Addressing these challenges places the research presented in this dissertation into the next frontier of machine learning research and application. Examining the feasibility in applying computer science to the domain of PRT videos is a novel undertaking and presents the opportunity to examine conditions that are important for expanding the capabilities of new technologies. These new technologies need to be trained to generalize tasks, looking beyond detecting specific activities into identifying key behaviors. The technologies need to be robust and implementable on ubiquitous devices under predictable environments. Most importantly, the technologies need to be diverse, supporting a wide range of variation in activities and individuals.

Continuing the research in automated metric extraction needs to examine different learning methodologies, particularly how deep learning can be incorporated. To accommodate this the dataset needs to be expanded. Creating the ideal dataset will focus on increasing the number of participants, and incorporate more variety regarding the people, places, and activities that are depicted. This will aid in the creation of algorithm that can generalize learning tasks to accommodate the naturalistic implementation at the center of PRT.

Extracting implementation metrics and diagnostic data is an important task the automated system can perform that alleviates the workload for expert clinicians. Examining the current formal assessments that are used in autism-related research provides an avenue for additional tasks that could be automated to the benefit and support of academic communities.

The data extraction algorithms and the prototype UI are intended to be a part of a larger system. The ideal design and implementation of this system requires additional

debates weighing the pros and cons of different system architectures. Distributed computing and low-power processing are two particular paradigms that warrant discussion. Additionally, the feedback modalities of the system need to be examined to determine the best methodology for relaying information to participants and utilizing performance metrics to motivate continued usage. In tangent with this, new technologies are emerging that could be incorporated into PRT and the automated feedback system to provide more flexible tools for parents to engage their children, and for clinicians to extract new data metrics.

The final thoughts of the dissertation briefly explore other domains that could benefit from the approach that has been presented. Educational and clinical environments are the most likely candidates as these reflect similar circumstances to PRT. Going beyond these domains, other recordings of dyadic and multi-person interactions could benefit from the dissertation research. Interview scenarios and police body camera recordings will be briefly discussed.

## 8.1 Alternate Approaches

### *8.1.1 Application of Deep Learning Algorithms*

Deep learning algorithms require a large amount of data to be successful. The size of the PRT dataset limited the feasibility of training deep learning algorithms from the ground up. Exploratory attempts to train a PyTorch DNN classifier did not achieve above an average of 60% training accuracy across the validation sets. With more data, deep learning approaches should be explored for their classification potential and the flexibility they could provide to addressing the problem. In particular, the multimodal aspects of the project could benefit from deep learning algorithms. This includes looking at different feature and decision fusion techniques that could be explored.

Multi-task learning has been used in similar domains as multimodal learning algorithms. Emotion and affect recognition are a common topic. Bidirectional RNN (Le et al., 2017), DNN (Parthasarathy and Busso, 2017), and deep belief network (Xia and Liu, 2017) have been utilized recently in order to classify arousal and valence from audio data. These approaches classify arousal and valence separately, then use the two values to determine affective state. Arousal is typified by level of emotion or excitement in a response, while valence measures whether the response is positive (happy, excited) or negative (angry, frustrated).

Thanda and Venkatesan (2017) incorporated video data into an automatic speech recognition network in order to capitalize on the visual cues when audio signals were insufficient due to noise or interference. They built upon existing research to design a DNN-HMM ensemble network. They validated their approach against a single task trained DNN model and concluded that the multi-task approach reduce word recognition errors when noise was present in the audio data.

Multimodal data is an important aspect of this project. The correlation between the audio data and the visual data in regards to attention was discussed in Chapter 6. Exploring multi-task learning solutions could be a useful method for extracting features that could be diagnostic for both visual attention recognition and VAD/speaker separation. As with other deep learning approaches, there was insufficient data to effectively train a multi-task classify. Attempts exhibited poor ability to converge, only achieving an average accuracy of 53%.

### *8.1.2 Self Supervised Learning*

Self supervised learning is a promising method that has the potential to aid in multimodal assessment of PRT videos. The audio-video action recognition research conducted by Owens and Efros (2018) could be applicable to this problem. This work trains a deep

neural network to learn the association between audio signals and actions in videos based on their temporal alignment. The synchronization of audio and action is important in PRT fidelity measurements. A fundamental goal of this dissertation is the identification of when a proper 'opportunity to respond' has been created by the interventionist. This involves the temporal alignment of the interventionist's instruction with the visual inference of the attention state of the child. Utilizing a training methodology similar to Owens and Efros's could make the recognition of this event easier to detect, and not require directly classifying the attention state. This would need to be explored based on its ability to generalize to difference sequences of actions that would reflect the variation in activities that can be anticipated in PRT videos.

Contingency is an important part of PRT implementation fidelity that has not been explored previously. This involves the interventionist relinquishing control of the motivational activity or object to the recipient quickly after a recipient provides an adequate answer. Self supervised learning techniques could be used to assess when this transfer has occurred and aid in the localization of preceding and succeeding vocalization.

The networks trained by Owens and Efros appear to associate the audio signals with alterations in the pixel data of the images to identify the audio's likely source. While the examples that have been presented show this to be fairly robust, PRT video probes would offer a challenging implementation environment. In particular, the toy noises and sounds from other activities may be difficult to localize if other motions are occurring simultaneously in the videos. An example from the publication illustrates the algorithm is capable of discerning speech from actors out of the frame. It would be interesting to explore if this translates to in-frame occlusion scenarios.

### *8.1.3 Alternate Sample Labeling and Attention Models*

The model of attention utilized for labeling the dataset was based on available literature regarding visual cues (Koegel, 1988; Suhrheinrich et al., 2011). The primary visual cues that the child is paying attention to the adult focus on the child's gaze, body orientation, and rate of movement. According to the resources, the child is attentive when looking directly at the adult or looking at an object in the adult's control. This is often accompanied by the child having his or her body oriented toward the adult, and the child may be reaching or pointing at an object in the adult's control. Generally, the child will not be rapidly moving during these periods of attention. Shared attention is a more prolonged attentive state. This is reflected by the adult and child engaging in a joint activity or sharing a visual focus on a mutually controlled object.

In contrast to the attentive state, the inattentive state occurs when the child is engaged in a solitary activity, ambulatory, or engaged in stereotypical or self-stimulating behaviors. This also includes disruptive behaviors such as acting out or having a tantrum.

For the dataset labeling, the individual one second segments were viewed and assigned a class based on these visual cues and the activity presented in the clip. In future research, alternative sample labels and models could be explored and compared in order to determine the most effective methodology.

Labeling the dataset for attention was based on if the child was attentive to the parent, inattentive, or engaged in a shared activity with the parent. However, this labeling simplification may have exacerbated the epistemic uncertainty condition in the dataset. As has been discussed in the previous chapters, circumstances depicted in the validation sets did not have sufficiently similar samples in the training set for accurate classification. Labeling the dataset based on visual cues, instead of attention state, may have aided in the generalization between samples.

Recommendations from the literature regarding the visual cues for determining attention were followed during the dataset labeling. This alternatively could have been accomplished by labeling the cues themselves, then inferring attention after classification. For instance, this would have included data labels regarding the child looking at the parent's face, looking at the parent's hands, engaged in solo play, or looking at a shared object. These labels could have been general enough to not require specific activity detection.

This approach may not be ideal considering the size of the dataset. Having more classes would dilute representation and exaggerate class imbalance problems. Using visual cues may have provided more opportunities for using pre-trained models and external datasets. While the attention labels were specialized to the problem, other dyadic datasets may be applicable to classifying visual cues.

Apart from visual-based methods, audio and biometric media could be utilized to assess engagement and attentive state. Analyzing the child's utterances in regard to the parent's instructions could be utilized to infer if the child was paying attention. Presumably, the proximity of the child's vocalization, the relevance to the instructions, and the perceivable emotion in the response could indicate if the child was paying attention when the instruction was initiated. The primary limitation of this approach would be in assessing individuals with developing communications skills that only demonstrate a limited vocal range. In particular, if the child is non-verbal, there may not be a vocalization to indicate attention. Using wearable sensors that monitor an individual's body metrics could also be used to infer attention. In particular, electrodermal activity and heart rate sensors could be utilized to detect engagement, arousal, or anxiety in the individual. If the child was wearing the sensor, the data regarding these states may be able to provide insight into the child's attention state. This would also be more robust to individuals that do not demonstrate stereotypical visual cues for attention.

The aforementioned approaches to modeling attention employed absolute classes for labeling. An alternative would be to treat labels as continuous instead of discrete. Discerning attention state, and similar human behaviors, can be subjective, especially when considering children with autism may not conform to stereotypical signs of attention. Instead of assigning a specific label, using a likelihood or probability value may make detection more accurate. This could be implemented in a similar fashion to emotion detection, where arousal and valence scores are computed on a continuum and the values are used to infer the subject's motivational state. A limiting consideration is how much data would be needed to train a sufficient model.

### *8.1.4 Sequence Recognition*

The research presented in the preceding chapters have followed the idea that most action recognition focuses on particularly diagnostic frames for classification, and not on an entire sequence (Schindler and Van Gool, 2008). While this has worked well for action recognition that is decidedly objective, the same may not be true for behavior recognition. In conjunction with the visual cue labeling, examining the sequence of frames may provide more diagnostic information for attention classification. This would provide flexibility in how the temporal information can be incorporated into the classifier. Feature fusion techniques that were described in Chapter 5 could provide a foundation of methods of merging the data when training classification models. These approaches could be compared to sequence specific classification techniques such as HMM and conditional random fields.

Detecting shared attention, as seen in the results from Chapters 3, 5, and 6, is particularly difficult, and largely task dependent. Unlike the attention state that has clear visual cues of attention, shared attention could have a more varied visual profile. Different activities could have periods where the child shows attentive visual cues, such as looking

at the parent or a shared object, or inattentive cues such as concentrating on an object in his or her possession during a game. The assumption in using the ST graph approach for visual cues relied on the classifiers recognizing proximity and shared gaze as indicators of shared attention. Research moving forward needs to address the sequence of events that indicate shared attention more overtly.

## 8.2 Ideal Dataset

Examining the attention classification shows that the variety of circumstances and actions depicted in the videos causes a high level of epistemic uncertainty in the model. Increasing the labeled samples would address this uncertainty by providing more diverse training samples that would help the model generalizing the classification tasks. Constructing a new dataset affords the opportunity to look at a holistic approach to applying artificial intelligence to PRT video probes. The ideal dataset for PRT should be designed to incorporate related audio-visual classification objectives in addition to the labels for attention and vocal utterance that are the focus of the proposed project. This would aid in attracting more research attention to technology-supported PRT, while providing a challenging 'in-the-wild' dataset for testing various algorithms. This dataset would need to be created in ways that would reflect the intended implementation environment and incorporate experts in the data labeling.

### *8.2.1 Leveraging Existing Video*

An astonishing amount of video is publicly available on popular sharing sites, predominately YouTube[1] and Vimeo[2]. With hosted videos numbering in the billions, it is a logical assumption that a large number of these videos will reflect dyadic interactions under similar conditions to PRT videos probes, with many direct samples of PRT. While

[1]https://www.youtube.com
[2]https://vimeo.com

the volume of videos presents a boon to finding data, it also makes the task of identifying acceptable samples difficult. While querying is the best method of finding candidates, it still produces a large quantity of results, making manual video selection time consuming. This also requires that the video descriptions adequately describe the contents. To effectively utilize YouTube and other sharing sites for samples, either teams of individuals are needed to review candidates or an automated evaluation process needs to be enacted. This could be implemented using the base information regarding expectations and assumptions in PRT scenarios that have been analyzed for this research. In particular, examining extracted information on the individuals in the frame using OpenPose could provide an opportunity to evaluate if the video represents an applicable scenario without manual intervention. Additional videos for analysis could also be obtained from autism research and resource centers; however, unlike publicly shared data, these videos are collected under privacy agreements which will limit their usage.

Along with the challenges to finding and selecting appropriate videos, using the videos in supervised algorithms will require manual labeling. Ideally, for scenarios involving human behavior, an expert should be used for sample labeling. This would provide the best opportunity for a machine learning model to correctly learn the application. In particular, the expert would be more adept at detecting subtle signs of attention, especially in individuals who may not exhibit the stereotypical visual cues.

Even without labeling, the additional video data could be used to improve classification models. The data could be included in training early layers in deep learning, giving the model more data for detecting low level features. The labeled data would then be used to train the final layers to detect the specific class. Similarly, the unlabeled data could be used to train an autoencoder that would aid in reducing the dimensionality of the input vector, which may improve classifier performance.

### *8.2.2 Behaviors and Activity Recognition*

Recognizing activities and making inferences based on individual and dyadic behavior are an integral part of PRT evaluation, as explained in Chapter 2. Of foremost importance to the research presented in this dissertation is the evaluation of the child's attention state. Expanding the dataset would not only benefit from more labeled samples of joint attention, but would be improved by adding labels regarding affect and engagement, as well as activity recognition. Addressing each of these tasks separately would allow for additional metrics to be extracted from the videos, providing more information that could be used for assessments. Additionally, as the concepts are influenced by one another, having additional labels would allow for joint model and multi-task training that could improve classification performance.

The attention state of the child, particularly social attention, is an important part of autism research and a fundamental aspect of the proposed project. Numerous studies have focused on analyzing and developing joint attention in children with autism (Jones and Feeley, 2009; Kasari et al., 2015; Lawton and Kasari, 2012). Having a labeled dataset on attention would aid this research by providing the opportunity to develop detection and classification technologies that could aid clinicians and parents in evaluating intervention procedures.

A dataset for joint attention was created using a wearable camera that provided the perspective of the child (Frank et al., 2013; Pusiol et al., 2014). This dataset was labeled in regards to the periods where the child was visually focused on the caregiver and performed an action that he or she was engaged with the adult. The MMDB (Rehg et al., 2014) dataset provides multiple camera views of interactions between a child and a clinician. These interactions are labeled on a five-class scale based on the visual cues that show the child is engaged with the clinician or a joint task; however, publications using the dataset for automated classification have combined labels to create few classes

(Rajagopalan et al., 2016, 2015; Tsatsoulis et al., 2016). This dataset was created using predetermined activities under laboratory conditions.

The proposed dataset would differ from these by providing a third person perspective of interactions in an uncontrolled setting. The attention labeling would focus on three states based on the child's attention: attentive, inattentive, and shared attention. Attentive refers to when the child is actively focused on the parent. In a PRT session, this is illustrated by the child being still, visually focused on the parent, and possibly reaching toward the parent. This likely occurs when the parent has control of the object or activity the child is wishing to engage with. The inattentive state would be indicated by the child engaging in solo activities, being ambulatory, exhibiting frustration or tantrums, or self-stimulating. Shared attention occurs when the parent is integrated into an activity with the child in a way that allows him or her to pause to prompt a learning opportunity without disrupting the activity. In this case, both attentive and shared attention states are likely indicative of joint attention or social engagement as defined in the datasets described above.

Judging affect or engagement has been a focus area for several multimodal classification research publications (Bosch et al., 2015; Castellano et al., 2012; Grafsgaard et al., 2014; Le et al., 2017; Parthasarathy and Busso, 2017; Rudovic et al., 2018; Xia and Liu, 2017). Approaches based on audio-video data are most applicable to the proposed project; however, adding physiological sensors, particularly electrodermal activity monitors, would be beneficial for affect and engagement studies. The aforementioned works, with the exception of Rudovic et al. (2018), did not focus on individuals with autism. Additional samples for children with autism would benefit the research community. These works also focused on engagement in interactions between applications or robots, while the proposed dataset would present scenarios for affect based on social interactions. Additionally, detecting affect of both the parent and the child has

been an important part of assessing PRT fidelity in research studies (Johnson et al., 2011; Kazdin and Whitley, 2003; Verschuur et al., 2019).

Like attention, labeling the dataset for affect should be conducted by two behavior experts and consist of labels for the parent and child. The labeling should follow the standard presented in other publications, focusing on levels of arousal and valence exhibited by the individual.

Although the approach to the current project has been to extract behavioral trends without regard to specific activity, the dataset would provide an opportunity for activity recognition. In particular, these activities would largely include child play and dyadic interactions. Dyadic interaction datasets have focused on common two-person activities, such as handshakes, kissing, hugging, and high-fives (Patron-Perez et al., 2010; Ryoo and Aggarwal, 2010; Sener and Ikizler-Cinbis, 2015; Van Gemeren et al., 2016).

Annotating the dyadic actions would aid in developing more robust activity recognition algorithms, especially for child-preferred activities. This also provides a video activity dataset for children with autism that is not focused solely on self-stimulating activities; however, if a child did engage in self-stimulating activities, this dataset could capture the activity that preceded the behavior. This could help put self-stimulating behaviors into context to aid in identifying triggers. Activity labels should include a brief description of the activity, and whether it is engaged in by the parent, the child, or both.

### *8.2.3 Object Detection and Tracking*

Object tracking in videos is an important step in evaluating PRT fidelity criteria involving parent recognition of a natural motivator. Often a child is motivated by one or more objects, such as toy cars or a cell phone. Parent identification and control of this object is important for engaging the child and gaining the child's attention in order to prompt a response. Outside of PRT research, this provides an opportunity for developing object

recognition, tracking, and saliency algorithms from ‘in-the-wild’ video data. In particular, the dataset could be useful for research regarding detecting or predicting objects that video subjects will interact with, and tracking objects under conditions with high rates of movement and occlusion.

Ideally, annotating objects in the video would provide coordinates for the bounding box on each frame. For PRT-based research, identifying the object that is currently the motivator along with who is in control (the parent or child) may be sufficient labelling.

One of the main areas caregivers struggle with is creating new instructions based off of the child’s object of interest. Having descriptions of the objects in the frame, such as color and function, provide the opportunity to research machine learning tasks directed at generating sample instructions based on visual information.

### *8.2.4 Vocal Utterance Detection*

The majority of the research and application of PRT has focused on improving social and communication skills. The approach relies upon effectively providing instructions, evaluating responses for validity and effort, and delivering reinforcement in a timely manner. Automatically evaluating these tasks is dependent on voice activity detection, speaker separation, and automated speech recognition. The ideal dataset needs to address these activities by providing labeling on audio data along with transcripts of speech and vocalizations.

As presented in Chapter 4, PRT videos offer a challenging scenario for audio recognition. As observed in the current video probe dataset, the audio contains not only the parent and child vocal utterances, but noises from toys and play activities, voices from electronic media, additional adult speech from session spectators, and general environmental noise. Research into extracting speech from noise often examines droning

repetitive noises from different locations (Drugman et al., 2016; Kim and Hahn, 2018). The noise in PRT videos is aperiodic and largely dependent on the activities in the videos.

Speech and vocal utterance patterns differ from other conditions. Parents often use child-directed speech patterns that are exemplified by elongated pronunciation of words, high voice pitch, and exaggerated excitement. This could make the speech more difficult to differentiate from noise or child speech. Child vocalizations may not consist of fully articulated speech, which may be more difficult to detect and separate from other audio.

For the ideal dataset, vocalization audio should be annotated with start and stop times and the speaker for each segment. Likewise, noises from activities or toys should be labeled based on their time interval and the object creating the sounds.

Going beyond voice activity detection and speaker separation, the vocal patterns observed in the PRT video probes would provide challenges for ASR technologies. In addition to the difficulty in identifying child-directed speech patterns from adults, recognizing the individual words could be problematic. In child-directed speech, pronunciation of words could be intentionally distorted. Depending on the ASR implementation and training dataset, this could lead to misrecognition of words from adult speech. Child vocalization and speech skills exhibited in PRT video runs the gamut from non-vocal to fully articulated, sentenced speech. The variation in speaker ability further compounds inherent difficulty in child ASR tasks. Using the corpora from PRT video probes would aid in creating robust ASR technologies and may be beneficial to researchers in the domain of speech pathology.

Studies involving lexical data would also benefit from labeled vocal data from PRT video probes. From a PRT perspective, analyzing the lexical data would aid in monitoring child vocal abilities in order to determine appropriate maintenance and target tasks to be utilized in training. The lexical data could also be used in conjunction with video, audio,

or physiological data to enhance affect research (Bertero et al., 2016; Mendels et al., 2017).

### *8.2.5 Contextual Information*

Information regarding the implementation context and attributes of the participants in the videos should be recorded and easily accessible. This should include descriptions of the environment the session occurred in along with demographic and skill information for the parent-child dyad. As addressed in Chapter 6, data on the child's age and his or her vocal skill could be useful information for researchers to improve classifiers (Rudovic et al., 2018), as well as provide a means for filtering the dataset to address specific configurations. The parent's familiarity with PRT would also be an important feature to note. Additionally, knowing if a video is a baseline, inter-treatment, or post-treatment assessment would be beneficial.

### *8.2.6 Dataset Logistics*

Collection of a dataset for PRT video probes should emphasize variation as much as possible. This variation would require classification algorithms working with the data to learn generalized representations for the target tasks. This generalization would be needed in order to account for the broad concept of PRT implementation.

The dataset used in the initial work for this project had seven participants in 14 videos with approximately 10 minute runtimes, giving a total dataset length of 140 minutes. Due to the variability in the domain, more participants and a greater number of videos should be collected. Due to task repetition in the video, the 10 minute runtime could likely be reduced to five minutes. Making considerations for participant recruitment and labeling time, 25 participants recording three videos each would likely be sufficient for continued research. This would bring the dataset total to 375 minutes of data.

Most of the research on PRT implementation has focused on children between the ages 24 – 60 months. The lower age limit for participants should be above the age when children begin to develop verbal communication skills. The upper bound is less significant, but likely should be restricted to 12 years of age if the dataset is expected to be directed at adult-child interactions. Ethnographically, the dataset should represent a diverse population.

Activities should be left to the participants to decide. Ideally, the video probes would be collected in a home environment. This would provide different environmental context along with variation in the activities across the videos. It is expected that the videos would largely contain dyadic play scenarios. It would be beneficial if the videos contained noncompliant behaviors such as tantrums and self-stimulation; however, how to illicit this data would need to be discussed between parents, clinicians, and researchers.

Ubiquitous handheld recording devices are the most ideal for capturing the PRT scenarios. This is due to their likely use in designing an application for PRT feedback. In addition to audio-video recording, discrete physiological sensors could also be utilized to collect electrodermal activity and heart-rate data for use in attention, affect, and engagement research.

Labels and annotations should be collected by two behavioral experts. This redundancy would be utilized to identify areas of agreement and disagreement in the dataset in order to identify scenarios that will likely be difficult for automated recognition. Final labels would include identifying video segments regarding attention, affect, and audio signals. A transcript should also be created identifying spoken language, noises from toys and activities, and all child vocalizations.

PRT specific data should also be collected on the parent implementation fidelity and the child's verbal communication skills. This should be conducted in a different manner than currently seen in ABA research and application. Ideally, evaluations should occur for

smaller increments of the video since one minute is a large amount of time from an automated system perspective.

## 8.3 Expanding Metrics and Assessments

Parent fidelity assessments and frequency of child functional utterances metrics, discussed in Chapter 2, are common evaluation criteria that are used in applied parent training as well as research studies. Examining the research surrounding PRT implementation and training presents other assessment methods. Table 8.1 presents the different assessment metrics that have been employed in the PRT literature along with speculation on the feasibility of employing automated data extraction. Examining these different assessments provides insight not only into how this approach could be used to reduce the manual cost of data collection in research, but also identifies additional information that could be collected and presented to the parent for more complete feedback.

Assessments given a low feasibility rating focus on self-reporting and examination of a history of behaviors. The Parent Sense of Competence Scale (Brookman-Frazee and Koegel, 2004) and Questionnaire on Resources and Stress (Holroyd, 1974) are concerned with the parents' perception of their abilities and current situation. This involves meta-cognition that cannot be inferred by data analysis. Aberrant Behavior Checklist (Aman et al., 1985), Children's Yale-Brown Obsessive-Compulsive Scales (Scahill et al., 1997), and Early Childhood Inventory v. 4 (Sprafkin et al., 2002) employ a questionnaire that asks for information on a variety of activities over a long period of time. Capturing all of the observation required for video analysis would require constant monitoring of the child, which is not feasible.

Several of the assessments would require specific classification models to be feasible for automated data analysis, or would require longitudinal cooperation. Griffith's Mental Development Scales (Griffiths, 1996) and Vineland Adaptive Behavior Scales (Sparrow

et al., 1989) would require activity recognition models for data classification. Parent Sense of Competence Scale (Johnston and Mash, 1989) is dependent on activities recognition and stress measurements which would require emotion detection models for analysis. Data from the Clinical Global Impressions-Improvement Scale (Guy, 1976) and Social Responsiveness Scale (Constantino, 2013) could be inferred from PRT video probes; however, this would require regular recordings over a period of several months to adequately cover the questionnaire materials.

Assessments that currently depend on video analysis, examine dyadic behaviors, or utilize specific language models would be the most likely candidates for data automation. The Autism Diagnostic Observation Schedule (Lord et al., 2000) and the Autism Diagnostic Interview (Lord et al., 1994) use video recordings of specific situations to evaluate the likelihood a child has autism. This could benefit from the automated data extraction and video clip creation described in previous chapters without requiring significant deviation from the current approach. The Mullen's Scale of Early Learning (Mullen and others, 1995) and Home Situations Questionnaire (Altepeter and Breen, 1989) are concerned with dyadic behaviors that could be recorded and processed in a similar manner to PRT video probes. The MacArthur-Bates Communicative Development Inventory (Fenson and others, 2007) and Preschool Language Scale (PLS-4) (Zimmerman et al., 2002) involve assessing demonstrated language skills. These measures could be automated by employing automated voice and speech classification techniques.

Table 8.1
Assessment standards and methods along with speculated automation feasibility.

| Assessment Standard | Evaluation Purpose | Method | Feasibility (High,Medium,Low) |
|---|---|---|---|
| Aberrant Behavior Checklist | Measures deviant behaviors including irritability, social withdrawal, stereotypic behavior, etc. | Questionnaire | Low |
| Autism Diagnostic Interview | Evaluates communication skills and reciprocal social interactions. Identify stereotypic behaviors. | Interview/ Observation | High |
| Autism Diagnostic Observation Schedule | Diagnoses autism based on four interactions | Observation/ Video | High |
| Children's Yale-Brown Obsessive-Compulsive Scales | Diagnoses if levels of fixations indicate the presence of a disorder. | Questionnaire | Low |
| Clinical Global Impressions-Improvement Scale | Generalized form for measuring effectiveness of treatments | Questionnaire | Medium |
| Early Childhood Inventory v. 4 | Diagnoses mental disorders including autism. Based on behaviors including selective mutism, eating/sleeping habits, problem behavior, etc | Questionnaire | Low |
| Griffith's Mental Development Scales | Evaluates child motor skills, adaptive behaviors, and coordination. | Observation/ Question-naire | Medium |

Table 8.1 Continued: Assessment standards and methods along with speculated automation feasibility.

| Assessment Standard | Evaluation Purpose | Method | Feasibility (High,Medium,Low) |
|---|---|---|---|
| Home Situations Questionnaire | Evaluates child compliance to instructions in different situations | Caregiver Question-naire | High |
| MacArthur-Bates Communicative Development Inventory | Measures vocabulary, comprehension, and language usage | Observation/ Question-naire | High |
| Mullen's Scale of Early Learning | Evaluates behavior and language between clinician administrator and child | Observation | High |
| Preschool Language Scale (PLS-4) | Measures child language comprehension, vocabulary, grammar, and inference. | Observation | High |
| Parent Sense of Competence Scale | Measures parent perceived confidence and skills when interacting with their child. | Questionnaire | Low |
| Observed Parent Confidence | Measures parent stress and confidence, and child affect, engagement, and responsiveness | Observation | Medium |
| Questionnaire on Resources and Stress | Measures emotional state of individuals in a household with a person with a disability | Questionnaire | Low |
| Social Responsiveness Scale | Evaluates behavior for autism diagnosis and the identification of social/communication deficits. | Questionnaire | Medium |

Table 8.1 Continued: Assessment standards and methods along with speculated automation feasibility.

| Assessment Standard | Evaluation Purpose | Method | Feasibility (High,Medium,Low) |
|---|---|---|---|
| Vineland Adaptive Behavior Scales | Evaluates response to verbal instruction, following direction, performing tasks, and problem behaviors. | Questionnaire | Medium |

## 8.4 System Implementation

Mobile technologies are an important consideration in designing the application architecture for the feedback system. Currently, video probes are recorded on handheld devices, most likely a smart mobile device. Due to the ubiquity of smartphones and tablets, providing a mobile interface is important as this would be a convenient way for users, the caregivers in particular, for accessing the system. The computer vision and audio classification algorithms have heavy computation requirements, especially if using OpenPose for visual pose extraction, as discussed in Chapter 5. The computation consideration means that two primary strategies can be employed for supporting a mobile interface, which are using a cloud-based distributed system and developing lower power classification processes.

### *8.4.1 Cloud Distributed Computing*

Distributed architectures afford the ability to disperse system functions and processing tasks amongst different hardware components and devices. Different network designs can be utilized to accomplish this. The most common designs are centralized, decentralized, and hybrid systems (Tanenbaum and Van Steen, 2007).

Centralized distributed systems are based on a client-server relationship between devices. The client device contains the front-end interface responsible for retrieving data

and displaying results. The server is responsible for the computation tasks on the input and generating the output. In a typical workflow, the client will collect an input, then signal the server, often transferring the data. The server performs the necessary processing, then returns a result to the client. Upon receiving the return message from the server, the client system will display the results.

In centralized systems, the client and server devices are specialized and arranged in a linear workflow. Decentralized systems provide a more generalized approach to a distributed system. In decentralized networks, each system is able to perform the same functionalities, leading to decentralized systems often being called peer-to-peer networks. This creates redundancy in the network that can be utilized to share processing tasks in order to balance the computation load over the network, as well as provide fail-safes if a given system ceases to function.

Hybrid systems seek to provide the control of having a central server, with the scalability of a decentralized peer-to-peer network. For this architecture, the client system connects with a server that acts as middleware, facilitating the distributing computation tasks amongst a larger backend network. This allows the network to be scaled with minimum updates, as only the middleware hub would need to know the new configuration.

Cloud-based distributed systems build these architectures in a fully online environment. This provides additional benefits to the system, including greater opportunities for remote access and increased redundancy.

Communication protocols between the system components can be conducted either synchronously or asynchronously. Synchronous communication requires the devices to maintain a connection for the entirety of the transaction. Doing so reduces the management requirements for processes by keeping individual transactions encapsulated. The major drawback is that maintaining this connection locks the client system processing thread until the transaction is complete, or a fail-safe time limit is exceeded. This could

lead to a poor user experience if the operator of the client system experiences a long delay. Asynchronous protocol addresses this by separating the communication into two separate calls. The first call initiated by the client system provides the request to the server. This call is terminated after the data transfer, allowing the client system to continue other processes. The second call occurs after the server has completed the computational task and performs a call back to the client system. This triggers an event process in the client system that will handle the incoming response.

The design of a PRT feedback system as discussed in the preceding chapters is dependent on a distributed system architecture. The expected use case is that the caregiver records a PRT session on a mobile device, then uploads the video to a central server via a client application on the device. Video processing and classification tasks would be undertaken in a server cloud that could leverage distributed processing in order to generate the analysis efficiently. This would be undertaken using asynchronous communication. The caregiver would be notified when the video automated analysis was complete, allowing him or her to view the results in the client application. This is also a necessary procedure for the human-in-the-loop components of the project. Uploading the video to a central server would make the video available to the clinician for review. The clinician's assessment could then be stored in the central network for access by the caregiver.

Utilizing this system architecture provides several benefits. Offloading the processing from the client device provides the opportunity to use more robust hardware for computational tasks. This ensures that regardless of the client device, processing times will be efficiently controlled by the central system. Additionally, multiple backend processing modes could be employed in parallel to further hasten processing. The video could be divided to allow multiple machines to process it simultaneously. An additional benefit is that software upgrades to the backend systems would be transparent to the client

applications, requiring no additional actions from caregivers. This would allow the system to continuously improve the classification models.

Costs, connections, and privacy concerns are the principal drawbacks of using a cloud-distributed computing paradigm. Operational costs related to the number of servers required are a logistical concern for the system. Efficient video processing would require sufficiently powerful servers to produce results in a timely manner. Offloading the processing to a cloud-based server also requires the user to maintain a network connection. This could be problematic for rural users with less reliable access to high speed networks. Additionally, as the data is uploaded into the network data management, security and privacy become important concerns. Many parents may be apprehensive of transferring videos of their children into the network.

### *8.4.2 Mobile and Low Power Computing*

Examining mobile and low power computing approaches can be a way to address the concerns regarding a cloud-distributed solution. Low power computing addresses optimizing complex computational processes for embedded systems and mobile devices. These types of devices often have limited processing capabilities and rely on battery power. Relevant competitions in low power computing have focused on various areas of computer vision, including object recognition, classification, and localization (Alyamkin et al., 2018; Debenedictis et al., 2016; Gauen et al., 2017; Lu, 2019; Lu et al., 2015).

PRT videos present a difficult application of low power computing (Heath et al., 2019e). The task requires making inferences regarding dyadic human behavior based on visual cues. This is a complex problem that may incorporate many sub-problems including human pose detection and body segmentation, gaze estimation, facial expression recognition, individual and dyadic activity detection, and engagement and attention detection. Additionally, due to the dataset, algorithms for approaching these sub-problems

need to be robust against occlusion, low-resolution, incomplete data, and low training set sample representation. This will likely involve taking a different approach than what has been described in previous chapters. The previous approaches relied upon OpenPose which is not currently conducive to low power scenarios. Successful implementation would need to discriminate vital keyframes to reduce the volume of data that would need to be processed. Additionally, the implementation would need to focus on key features of the individuals, such as face and hand location, instead of searching for every body and landmark point.

Employing the classification and analysis tasks directly on the recording device eliminates the need to transfer the data to an external source. Potentially, if the low power algorithms were efficient, this could reduce the processing time. However, it is expected at this current state that the computational system afforded from external hardware would compensate for the communication time. Eliminating the external transaction would alleviate security and privacy concerns. The data would not need to leave the caregivers device, making it less susceptible to nefarious internet activity.

The limitations of this approach are based on the performance of the employed device. Currently, mobile devices are limited by processing power and energy consumption, as opposed to a distributed solution where these resources are potentially unlimited. This constricts the algorithms and data processing methods that can be utilized. This also makes the process more device dependent. Essentially, an individual's experience will be dependent on the device he or she has available. This could be prohibitory to families that cannot afford a compatible device.

Removing the data transfer completely from the system would also restrict the interaction between the clinician and the caregiver. One of the goals of this project has been to examine how the system could facilitate this connection, and reduce the time required for clinicians to provide support. Without transferring the video, the clinician

will have less information for forming conclusions. A compromise based on decentralized networks could reduce privacy concerns while maintaining this connection. A peer-to-peer-based network and providing a communication link directly from a parent's system to a clinician's system, without centralized routing or data storage, could be utilized to mitigate concerns of unauthorized data access.

## 8.5 Feedback Modalities

Assessment and feedback are important for aiding individuals in acquiring new skills. The type of assessments and the methods for delivering feedback need to be considered to ensure they are conducted in an effective manner. For a PRT feedback system, interventionists need to be provided with data they can use to capitalize on opportunities to provide instruction, evaluate language uses, and ensure reinforcements are appropriate. Presenting this information can be done in real time, during the session, or off-line after the session has concluded. Each of these strategies offer different affordances and drawbacks.

Both systems are largely dependent on the preferences and unique circumstances of the individuals and the environmental context they have available. There are some areas of overlap, however, the approaches for this application are, for the most part, mutually exclusive. Ideally, the system would follow a person-centered paradigm allowing for the configuration of both approaches on a per user basis.

### *8.5.1 Off-Line Feedback*

In the PRT feedback system, there are three main types of data metric gathering opportunities to exploit for providing feedback for off-line feedback. The first is the data that can be extracted from a single video in isolation. The second examines the

progressive and aggregate data from multiple sessions over time. The third is data from clinician review of videos, usage, and data metrics.

Providing feedback to the parent needs to focus on how his or her adherence to PRT methodology, and the relationship between his or her actions and the response of the child. In the scope of this project, the primary focus is to understand when the child is attentive to the parent and when the child vocalizes. For this, classification models to detect the attentive state and identify the parent and child speech segments can be used. Feedback can then be provided on the relationships between the attentive state, the child vocalizations, and the adult speech. The expected correlations, and what can be inferred by comparing the baseline video probes to post-training video probes, is that child vocalizations are positively correlated with increased attention and negatively correlated with adult speech. This aligns with PRT methodology that states an effective instruction is clear, limited to the language level of the child, and delivered when the child is attending the parent. Additional metrics including the mean length of a child utterance, or a breakdown of child responses based on types of prompts, or spontaneous speak, could also be important feedback for the parent. Affective states have been noted as an important side effect of PRT training sessions. Future data collection could include inferring affective state from the video and audio (D'mello et al., 2008; Grawemeyer et al., 2016; Sanghvi et al., 2011) data as a means of easing parent stress levels (Lecavalier et al., 2017).

Feedback on individual videos also provides the opportunity to give corrective notifications based on the environment the session occurred in and the recording quality. This will aid in future data collection and assessment as well as reassure the users that they are using the system correctly.

In addition to providing metrics, the system should promote self-reflection on each session as part of the feedback system. Self-reflection is an important part of self-regulated learning and helps the learner understand his or her performance on a given

exercise. It also provides insight on how to continue to improve. During the self-reflection phase, the learner can compare performance to goals, build self-efficacy, and adapt the process for the future (Isaacson and Fujita, 2006).

The feedback system, like PRT itself, is intended to be utilized over a prolonged period of time. This provides the opportunity to examine both the child's and the parent's progress over time. Particularly, this is helpful in identifying plateau periods in the child's demonstration of skill that should elicit a change in instructions and the identification of new target and maintenance skills. Implementation of PRT is similar to a motor skill. Initially, large improvements will be made by the learner, but the rate of improvement will slow with the rate of mastery (Kitsantas and Kavussanu, 2011). At a point, feedback directly related to the parent will become less important, as there are fewer corrective actions they need to take to maintain PRT implementation fidelity. This is another reason why it is important to emphasize child data metrics, affective improvements, and self-reflection.

The role of the clinician in the system is to provide additional pointed feedback to the parent, help the parent understand the metrics that have been gathered, and to aid the parent in planning and executing PRT sessions. The planning portion of their role is particularly important during plateau periods when the parent needs to adapt PRT to incorporate new skills or activities. In addition to this, keeping the clinician involved facilitates a social obligation to continue using the system, and may aid in the confidence of the parent by maintaining a supportive connection between client and expert (Kitsantas and Kavussanu, 2011; Pintrich, 2000).

There are three related drawbacks to offline feedback. First, and probably most substantial, is that reviewing the feedback requires an additional time investment from the user. At the least, the parent will need to upload the media, wait for processing, then review the results. Gathering more metrics will increase the overall review time.

Improvement will then be dependent on the parent spending a sufficient amount of time on reviewing and understanding the feedback. Addressing this drawback requires ensuring that the system processes data quickly and that the UI design is intuitive. Additionally, the data needs to show a clear benefit for the parent in order to ensure compliance.

Second, feedback does not occur directly after an action. This delay could cause a disassociation with previous actions that could make correcting behavior more difficult. This can be rectified using the video records to review and self-reflect on the sessions in addition to receiving feedback(Suhrheinrich and Chan, 2017); however, this further increases the time commitment.

Third, in addition to not associating feedback to an erroneous action in the moment, not providing real-time assessment prevents the parent from immediately correcting his or her actions. Being able to immediately correct an incorrect action helps to display the contrast between appropriate and inappropriate actions. This will not only help the parent learning the correct behavior, but will also aid in fostering self-confidence in his or her abilities.

### *8.5.2 Real-Time Feedback*

Considering the discussion on feedback above, looking at effective real-time feedback is challenging. There are two primary aspects that need to be considered for providing feedback to the parents during PRT sessions. First, the important metrics, data, and feedback goals that are important in real-time interactions needs to be addressed. Second, the feedback delivery method needs to be examined.

Whereas offline feedback can look at collective data over one or more sessions, real-time feedback should focus on actions that can be immediately enacted or corrected. In observations of feedback sessions between parents and clinicians, real-time feedback was most often demonstrative. This largely consisted of modeling a type of instruction for

a given task, showing how to engage the child or elicit his attention when he was involved in solitary play. This is not totally conducive to an automated real-time system. In the observations, the clinician is capitalizing on the limited amount of time. Distracting the parent or child is not a primary concern. Automated systems may be best implemented by using information gathered at the beginning of a session to improve implementation toward the end of a session.

In PRT, the parent is expected to provide two opportunities for the child to respond per minute in the session. An automated feedback system could attempt to detect if these attempts have occurred, and if they have not, prompt the parent to engage the child. Similarly, variation of task and instruction is important. The system could track instruction type variation along with target skills to prompt the parent to differentiate.

Real-time feedback could be used to determine if the child is paying adequate attention to the parent. The prompt could initiate the parent to provide the instruction at an opportune moment; however, this could be problematic if the parent is focusing on catching the prompt and not directly observing their child. Similarly, in the moment assessment of the session context could help the parent identify the object the child is interested in.

In conversations with parents learning to implement PRT, one of the challenges that was expressed was composing appropriate instructions for a given task during the session. A more elaborate real-time automated intervention could analyze the object or activity the child is motivated in along with the child's language level to determine sample instructions for the parent to enact. This could also be utilized to address instruction and skill variation.

Real-time expert-based feedback was addressed by Machalicek et al. (2010). In this study, an ABA session involving a teacher and student was broadcast to a behavioral expert in order for the expert to provide feedback to the teacher on her implementation. Although the research conclude that this method was successful for training the teachers,

there were a couple of limitations that should be considered. The technology used for the tele-conference was a hindrance. Several of the teachers in the study had difficulty connecting the hardware and some students were distracted by its use. Additionally, the researchers report a decline in fidelity for some teachers that they theorized may have been due to the teacher anticipating negative feedback. This illustrates that the success of real-time feedback is likely to be dependent on the behavior and preferences of the participants.

For the feedback to have the most universal appeal, it should be implemented in the least intrusive means possible. Selecting a mode for the feedback delivery is a challenging design consideration. Overtly visual or audio clues are likely to be the most distracting; however, it would be difficult to deliver complex notification without the use of either medium. If the session occurs in a pre-designated area, then a discrete monitor could be placed in the room to notify the parent, particularly if the system provides example verbal instructions. This may still be distracting to the child, and pure PRT implementation should not be confined.

Haptics may provide a methodology for conveying covert notifications. Using a simple pattern-based system could prompt the parent to add a maintenance instruction, or that it has been too long since the parent last engaged the child.

Reviewing the drawbacks listed above, real-time feedback could be distracting depending on the individuals involved and the modality used to provide notifications. The notifications themselves need to be simple and easy for the parent to digest quickly to adequately capitalize on the prompt. Real-time assessment adds additional technological constraints in order to process the video stream quickly and provide results.

The final drawback of real-time feedback is that its worth to the user diminishes over time. Real-time feedback is best suited to correcting behavior. As the parent gains skill and efficacy in PRT, the necessity for correction or prompting is reduced. This would need

to be weighed along with the developmental complexity against the benefit to the parent in order to determine its true value to the system. Most likely this would be largely dependent on each individual case.

## 8.6 Adapting PRT to New Technologies

The primary focus of this dissertation has been how current multimodal data processing and machine learning strategies can be applied to aid in the support of parents implementing PRT. This has been conducted under the assumption that PRT is immutable. Looking at PRT and technology as co-adaptable provides opportunities to incorporate new technologies that are currently not being utilized. Examining these new technologies provides affordances that would benefit clinicians, parents, and children during PRT sessions. These technologies could provide opportunities to gather metrics that are not currently tracked, allow for the expansion of imagination and possibilities for participants during sessions, and aid in automatically detecting activity and attention. In particular, wearable sensors, additional recording, and augmented reality (AR) are worth a brief discussion.

### *8.6.1 Biometric and Inertial Sensors*

Wearable sensors facilitate an opportunity to retrieve information directly from the users. For PRT video probe analysis, this supplements the video data by providing information about each individual directly. This can be useful in activity recognition using inertial sensors such as accelerometers and gyroscopes. Biometric sensors, such as electrodermal, heart-rate, and brain activity monitors, can provide insight on each individual's affective state, especially when it is not apparent in the video or audio recordings.

Accelerometers and gyroscopes have been studied as wearable sensors for human activity recognition (Attal et al., 2015; Casale et al., 2011; Zeng et al., 2018). For

individuals with autism, the focus of inertial sensors has been detecting self-stimulating behavior (Coronato et al., 2014; Kientz et al., 2007; Rad and Furlanello, 2016). Only relying on these types of sensors would not be adequate for PRT analysis. The referenced studies focus on activities that have repetitive periods. Detecting attention has some dependence on movement; however, it also requires more spatial information to make an adequate inference. Additionally, data streams from the sensors would need to be synchronized across both individuals in order to provide a perspective on dyadic interactions. Despite these drawbacks, the data could be useful in conjunction with the video, particularly during periods of occlusion.

Electrodermal activity and heart-rate sensors have been studied in regard to engagement and emotion detection (Conati et al., 2003; Fletcher et al., 2010; Picard, 2009). Skin conductivity and heart rate have been associated with states of arousal and can be used to infer engagement, stress, or anxiety. Using these types of sensors can be an effective means of determining a wearer's affective state in regard to a given activity. A positive affect and improved stress has been listed as a benefit of parent-led PRT. Measuring the child and parent's affective states during PRT sessions, and progressively over time, could aid in motivating continued use of the treatment along with promoting greater self-efficacy in the parent regarding his or her ability to implement the interventions. These sensors would provide a more precise way of measuring efficacy than relying on self-reporting or making inferences through video or audio data. As suggested in Picard (2009), a child with autism may not visually or vocally express emotion. Using the sensors would provide a means to collect this information.

Obtaining the affective data regarding the child could also be used to facilitate a greater connection between the parent and child. Picard (2009) relates a use case scenario where a teacher could pair a device with a child's wrist sensor in order to better understand engagement in the material. A similar system could be beneficial for parents

during PRT. Providing haptic cues to the parent on the emotional state of the child during a PRT session could help them determine periods of engagement or frustration that would allow them to continue, adjust, or stop the treatment. Use of the affective information in conjunction with the video as part of part of the post-session review could help parents relate how their actions affected the emotional state of the child.

Electroencephalograms (EEG) have been utilized in making inferences of engagement and attention to tasks (Billeci et al., 2016; Li et al., 2017). In these publications, engagement is detected using EEG headsets by looking for spikes in brain activity. EEG is also becoming a commonly used instrument for early autism detection (Bölte et al., 2016). Using EEG in PRT sessions would be an additional method for inferring child engagement in an activity. In their study, Billeci et al. (2016) concluded that they could differentiate when a child was engaged with an activity and when he or she was not. The sessions were conducted in a 'semi-natural' setting, where the child interacted with a clinician. It would be interesting to see the results if the same experiments were conducted in a PRT session. In their scenario, the activities were likely selected by the clinician, or the child had a limited number of options. The PRT experiments would need to be conducted in the child's home environment, with access to the child's favored activities and objects. During PRT sessions, the child could be engaged in solitary play, or be attentive to parent. This research would address if these two states of engagement could be differentiated.

Implementing sensors as part of the data collection process for PRT sessions could provide insight that would otherwise be difficult to glean from only the video and audio modalities. This does come with the caveat that the additional technology would provide complexity for the users and could lead to technical difficulties during installation and use (Marcu et al., 2012). Additionally, physical activity can have an effect on each of these types of sensors, causing false readings (Sun et al., 2010). Utilization of the data might

require using video context to ensure it was properly analyzed. A further drawback could be the participants acceptance of the technology. Both the parent and the child may have an aversion to wearing a device. The device itself could be a distraction during the PRT session, which may result in non-compliance.

In addition to exploring wearable sensors, attaching sensors to objects used in PRT sessions could provide benefits for PRT and automated analysis. 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 (Dorfmüller, 1999). Utilizing common color patterns, similar to image codes (Mehner et al., 2015), 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 (Zhang et al., 2017b). 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 if they could be used for localized tracking. This could provide a cost-effective solution for attaching sensors to objects.

### *8.6.2 Recording Devices and Perspective*

In previous chapters, 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 affords 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 (Jazouli et al., 2016; Shahroudy et al., 2016; Zhao et al., 2017), improves occlusion handling (Pi et al., 2016), and aids in estimating visual focus (Wei et al., 2017). Having perspective in the image would help 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.

Different camera perspectives have been used in studies on joint attention and child engagement. Static exocentric cameras are most commonly used and can include multiple cameras capturing multiple angles of the interaction (Rajagopalan et al., 2016, 2015; Rehg et al., 2014). Using this configuration has several benefits. This configuration works well for laboratory conditions, but does not translate as well to real-world scenarios due to the equipment and installation required. In addition to multiple exocentric cameras, the Multimodal Dyadic Behavior dataset (MMDB) included an egocentric camera worn by a clinician during interactions with a child. Egocentric cameras for human activity or

behavior recognition are more common in the field of human-robot interaction (Foster et al., 2013; Li et al., 2012).

There are three concepts that are relevant to the use of video in the proposed system: detecting and analyzing child and parent behavior, providing multimedia material for clinician review, and promoting self-reflection in the parent. We can look at these under the consideration that the parent is wearing the camera.

For analyzing behavior and activity detection, the egocentric camera offers several affordances. It is known, relative to the scenario, where the camera is spatially located. This allows data processing algorithms to make assumptions when analyzing the interactions. Most important to the research that has been done thus far is the assumption that when an individual is looking at the camera, he or she is likely attending the person wearing or holding it. This eliminates the need for a machine learning algorithm to estimate visual attention and is likely more accurate than the current method used in the initial work discussed in Chapter 3. Additionally, the wearable camera is likely to be closer to the interaction than is depicted in exocentric video probes. This could provide the opportunity to use eye tracking algorithms to gain a further understanding of the subject in the frame's visual attention. This would prevent a few of the difficult scenarios that were observed in the video probes used for the initial work on the proposed projects. The system had difficulty determining when a child was looking down or looking at the camera. Also, camera perspective was an issue when the algorithm would incorrectly attribute the child's object of visual attention as the parent, when the parent was actually in the background of the frame behind a toy that was the true object.

Improved audio recording quality could be an additional benefit of egocentric cameras. The closer proximity of the recording device to the interaction could allow the device to capture lower energy vocalizations; however, the wearable microphone could also record additional noise, such as clothing rustling.

Having the camera as a wearable also eliminates the need for an additional person to operate the recording device. Ideally, attaching the camera to the parent would also make it easier to keep the camera directed at the child. As the parent is expected to continuously observe and engage the child, he or she should be oriented toward the child.

The optimal use for wearable cameras is restricted to certain types of interactions. In the MMDB dataset, both the child and the clinician are sitting across one another at a table. The activities they engage in are undertaken seated, maintaining distance between the two individuals. This ensures the child is kept in frame, and movement is minimal. This scenario does not account for the freedom of activity expected in PRT interventions. The activities in a PRT session could include physical movement that would affect the quality of the recording and the ability of the parent to keep the child in the frame. Similarly, the perspective of the parent and child facing each other cannot be guaranteed. For example, a video from the dataset examined in initial work on the project depicted a child sitting in the parent's lap watching a cell phone video. This is a valid PRT scenario, as the parent can prompt the child to vocalize by pausing the video and issuing an instruction. If the parent was wearing a camera, the camera would capture the child from behind. Additionally, the close proximity to the parent and child would limit the field of view of the camera, resulting in a potential clip that cannot be processed by the analysis algorithms, limiting the value for clinician feedback.

Ruminating on how the clinician would use egocentric perspective illustrates similar benefits and challenges to the automated analysis. The clinician would be presented with video that focused on the child's behavior but may be left to assume or infer the action taken by the parent. This may be particularly problematic with parents in the early stages of learning that may not be efficiently following PRT methodology. In these cases, without direct knowledge of the parent's actions in the video, it may be difficult for the clinician to provide specific feedback.

The effects of using egocentric video for self-reflection from a meta-cognitive perspective is unclear. In accordance with self-regulated learning methodology, the video would meet the criteria of self-monitoring; however, it is my opinion that part of the importance of reviewing video tape for performance is experiencing the event from a different vantage point. This allows the individual to understand how his or her actions are situated in the complete interaction. Only providing an egocentric viewpoint would allow the parent to replay the interaction, but would largely lack additional information that could be useful in reflecting on his or her performance.

Attaching the camera to the child to gain his or her perspective has been explored by Marcu et al. (2012) and Pusiol et al. (2014). Pusiol et al. found mixed results in terms of identifying periods of joint attention using a camera attached to the child's head, ultimately concluding that parameters such as object movement during interactions is more diagnostic of attention than periods where the caregivers face was in the field-of-view of the camera. This would also likely be the case in the context of the PRT video probes. During periods of inattention, the parent's position in the child's field of view may not be a diagnostic feature.

Regarding clinician feedback and parent reflection, this video perspective becomes more powerful. Providing the child's point of view of the interaction could offer insight into his or her experience during the interaction. This could help from an empathetic standpoint, allowing the parent or clinician to reflect on the child's perspective of the treatment. This could illustrate periods of joy and frustration in a way that would help the adults tailor future interactions. The importance of this perspective was expressed by the parent participants in Marcu et al. (2012).

In regard to the video sound track, speaker separation was described as a challenge facing automated audio analysis in the video probes. Under the current assumption, this would be audio 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 (Pawar et al., 2017; Shivakumar et al., 2017; Xu et al., 2009, 2008) employed discreet wearable microphones for the child. Using this approach on both of the individuals in the interaction would enable 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.

### *8.6.3 Augmented Reality*

Augmented reality (AR) is a methodology for projecting virtual images into real-world space. Using AR creates an 'intuitive metaphor' and affords the opportunity to seamlessly blend interaction between the physical and virtual contexts (Billinghurst and Kato, 2002). For educational systems, this provides a means of easily displaying the effect of manipulating objects. For example, an AR system is presented along with a sandbox that can be utilized to explain topographical features and aid students in learning to read a topographical map (Beals, 2017). This system allows students to construct mounds and troughs in the sand and observe how this affects the overlaid topographical map. Through this interaction, they are able to visualize how their actions influence the overlay being projected. This helps situate the learning objective in context, in this case interpreting topographic symbolism, that can help solidify abstract concepts (Shin et al., 2016).

AR research for children has focused on presenting information in a new and interactive medium. These studies report favorable results in terms of child enjoyment and teacher satisfaction (Lim and Park, 2011; Rasalingam et al., 2014), even extending to preschool aged children (Yilmaz, 2016). Studies regarding children with autism have focused on imaginative play (Bai et al., 2015), selective and sustained attention (Escobedo et al., 2014), and improved social etiquette (Liu et al., 2017).

Imaginative, or symbolic, play is the child's ability to use his or her imagination to transform objects and environments as part of an intrinsic narrative scenario. This is often cited in literature regarding autism as being linked to underdeveloped social and language skills (Orr and Geva, 2015; Stahmer, 1995). In Bai et al. (2015), the authors projected a virtual overlay onto a physical block to change its appearance to a car. They concluded that after using the AR system, the children showed a greater propensity for symbolic play without the use of the AR. While symbolic play is not a specific area being addressed in PRT methodology, incorporating AR technology would easily allow for the incorporation of imagination into intervention sessions aimed at other target behaviors

The use of AR has been examined in DTT. Due to the inherent repetition in the tasks, the recipient may lose attention and become noncompliant. More generally, deficiencies in selective attention, the ability to focus on a single activity, and sustained attention, the ability to maintain one's focus on a single task for a prolonged period of time, are common in children with autism. Escobedo et al. (2014) used AR as a means of promoting engagement for the children during the treatment sessions. That was accomplished by using a mobile phone to display an overlay on learning materials. They found that this method resulted in increased selective and sustained attention for their participants.

Utilizing BrainPower, a commercially available AR system developed for individuals with autism, Liu et al. (2017) explored how headset-based AR could aid individuals in improving social interactions. Their system used facial detection algorithms to identify faces in the wearer's field of view and project an animated overlay on the person's face. The study reports that this led to a caregiver-reported increase in eye contact and other social skills. Additionally, the authors state that measures of common negative effects associated with individuals with autism, such as irritability and social withdrawal, along with stereotypical behaviors, were greatly reduced after using BrainPower.

These studies produce interesting results, however, there are some key caveats. Primarily, the number of participants and the number of sessions is limited. This limits the generalizability of the results, and does not address the fact that the conclusions are the result of a novelty effect due to the new technology. This novelty effect could also be problematic for individuals with high levels of idiosyncratic behavior. In particular, the research presented by Liu et al. (2017) provokes skepticism. The authors of this work only used two individuals for their case study, and the authors of the article have a financial stake in providing exemplary results. Regardless of these caveats it does show the potential of AR technology for children with autism.

Including AR in parent-led PRT sessions could benefit from the development of symbolic play associated with AR research. Selective and sustained attention are less likely to be an issue in PRT based on the requirement that the child select an activity that is naturally motivating. Implementation of AR in a PRT session could be undertaken in several different ways. This will be discussed for both collaborative AR environments and displays restricted to the parent or clinician. Providing displays for the child only has its benefits, but since this would limit the participation of the parent, it will not be discussed in detail.

Incorporation of a collaborative AR environment for PRT sessions would provide more opportunities to engage the child by providing a configurable environment. The potential to virtually augment the appearance of objects and environments would increase the number of activities that could be utilized. This could be particularly powerful for helping children generalize their verbal skills to new tasks. For instance, a simple example would be if the overlay on a wooden block transforms it into a yellow car. The parent could issue an instruction for the child to identify the color of the object. Changing the AR to transform the car into a submarine would provide the opportunity to again ask for

the color of the object, reinforcing in the child that the color is not dependent on the object representation.

Different overlays in the system could also aid the child in progressing to written communication skills. The AR technology would allow for providing text descriptions of objects that would aid the child in reading skills. Additionally, parent instructions could be displayed in text as well as spoken verbally.

It would be interesting to research how the technology could be used to facilitate the parent taking control of the activity to introduce an instruction. This could potentially occur in one or two scenarios. The first scenario involves an object within the augmented experience being the motivator for the child. For this, the parent would need to physically exude control over the object, in the same manner as non-AR sessions. In this case, the AR is supplementary to the interaction and serves primarily to enhance the experience of interaction with that object.

In the second scenario, the AR experience is central to the child's motivation. In these instances, the parent can exude control over the virtual environment to gain the child's attention to issue an instruction. Ideally, the parent would be fully integrated into the activity the child is participating in. This would afford the option of working the control into the endogenous fantasy of the activity, making the learning objects more integrated. For example, instead of the parent physically taking control of an object from the child, the parent could initiate an overlay change on the object that would alert the child to an instruction.

An additional research question for AR implementation would be how engaging the parents find AR activities compared to traditional PRT implementation. If the parents are more engaged in the interaction, they would achieve a greater affective boost from participating, and would be more likely to continue the sessions.

There are two technical implementations for creating collaborative AR environments based on the technology being used. These approaches can use linked individual displays, such as mobile phones, head-sets (Google Glass, Microsoft HoloLens), or shared displays, primarily consisting of projectors. Each of the technologies would have affordances and disadvantages.

A dyadic AR experience was facilitated in a study conducted by Dierker et al. (2011). This system allowed two individuals in the same room to view virtual exhibits projected onto a physical museum. This allows the individuals to collaborate on the placement of the exhibits. This is a similar scenario that we could see being enacted for PRT sessions. In the sessions, the parent and child wear separate headsets and interact with a shared set of objects. This would require two synchronized displays; however, it is likely a simpler implementation than projector-based solutions. Having the two headsets would afford the opportunity to record both participants' perspectives of the interactions that could be used for self-reflection or clinician review.

Shared display systems could be implemented using a single screen device shared between the two participants, or a projector creating an environmental AR overlay. Using a projector likely causes a greater sense of immersion compared to screen-based implementations due to the removal of the screen as a conduit for the experience. Using the projector does create additional technological challenges. The virtual image would be more dependent on the surface it was projected upon. This would likely place a requirement on the types of spaces the technology could be used in. Projected images also have the challenge of ensuring each individual experiences of the same AR display. A participant's view of the virtual image could be influenced by his or her perspective of projection (Benko et al., 2014). Depending on the distance between individuals in the interaction, and their orientation, multiple projections may be needed to facilitate a simultaneous AR experience. Considering this, the approach undertaken in the sandbox

video (Beals, 2017) with a top-down projection is ideal configuration for collaborative projector-based AR using the current state of technology.

Parent-only AR systems would focus on delivering real-time feedback in an unobtrusive manner. This could largely consist of environmental and object overlays that would help the parent improve the session context, both for data capture and for instruction delivery (Liu et al., 2017; McMahon et al., 2015). For instance, the headset could relate information on environmental clutter or excessive noise so that the parent could take corrective action during the session. Similarly, the display could be linked to an exocentric perspective camera and inform the parent when there were substandard conditions for capturing the parent's or the child's body pose or vocal data.

Utilizing object recognition, or detecting objects with QR tags (Yilmaz, 2016) along with contextual information regarding the child's vocal ability level, could provide the opportunity for the system to display sample instructions. The system could recognize the child's natural motivator, then generate a variety of instructions based on this object or activity. This would aid in generalizability of PRT to new activities, and help the parent ensure they are delivering the appropriate amount of maintenance and target skill opportunities.

A novel concept for using AR could involve incorporating the clinician into the session via telepresence (Pejsa et al., 2016). Telepresence essentially is an AR video call, where one individual is projected into the environment of another. Using this technology could virtually situate the clinician in the PRT session through the parent's AR display. This would help the clinician provide in-context feedback while the parent is interacting with the child. Additionally, the parent's session could be projected into the room of the clinician to give them a full view of the interactions.

## 8.7 Other Applications for Approach

### *8.7.1 Diagnosing Autism*

Diagnosing autism through computer vision and speech analysis has received attention from the research community. Computer vision techniques have focused on behavior analysis (Hashemi et al., 2012; Martin et al., 2018; Rajagopalan et al., 2013), visual focus (Alie et al., 2011; Jiang and Zhao, 2017), and neuroimaging (El-Baz et al., 2007; Elnakib et al., 2011). Vocal-based diagnostic systems have focused on speech patterns (Xu et al., 2008) and sentiment analysis (Marchi et al., 2015). The intent of these approaches is to find an effective method of identifying individuals with autism to make diagnoses more accessible. Essentially, the goal is to remove the human component of analysis to reduce the cost of evaluation.

The classification methods described in this dissertation are similar to the works above, but differ in the methodology and target analysis tasks. Computer vision techniques focused on behavior are often looking for specific actions such as stereotypical behaviors (Rajagopalan et al., 2013) or head movements (Martin et al., 2018). The goal of the research presented in this dissertation has been to examine attention in a general sense, divorced from specific actions. In regards to audio data, the works mentioned in the previous paragraph (Marchi et al., 2015; Xu et al., 2008) assume the participant is vocal. As discussed in Chapter 4, individuals with autism may exhibit limited vocal skills but produce non-speech vocalizations.

The NODA application (Nazneen et al., 2017, 2015; Solutions, 2018b) introduced in Chapter 7, in conjunction with the diagnostic assessments presented in Table 8.1, provide a scenario where the methodology discussed in this dissertation can be applied to diagnostic tasks. The NODA approach is to direct the parents to enact scenarios with their child while creating a video recording. The recording is then manually analyzed by

behavioral experts to create a diagnosis. Like the PRT video probe evaluations, the methodology described in the preceding chapters could be utilized to create analytics and key video segments that would aid clinicians in their assessment. Keeping the clinician as part of the diagnostic process, in contrast to the approaches above, would allow for more versatile feedback and immediate treatment options.

### *8.7.2 Classrooms and Educational Environments*

The approach for detecting attention presented in this dissertation could be applied to classroom situations. Although classrooms contain multiple individuals, during a teacher's lecture period each student should be focused on the teacher. This is essentially a dyadic interaction similar to what has been discussed in previous chapters. The expectation is that each student is attentive to the teacher, reflecting similar visual cues as were addressed in assessing PRT fidelity. In particular, these would include visual focus on the teacher and limited interaction with other students in the class. Visual focus is the main methodology for detecting attention that has been employed in research (D'Mello, 2016; Raca et al., 2015). This is likely to be the most overt visual cue of engagement; however, the approach becomes more difficult in scenarios that would require discerning note taking from a student having his or her head bowed or resting on the desk.

Implementation of the system could address different aspects of the classroom. Teacher and coursework evaluation could be primary use cases. Gathering attention metrics along with verbal requests and responses would be expressive of activities or subjects that were engaging. This would aid teachers and teaching instructors in assessing a particular teacher's classroom demeanor. Looking at the data, especially if multiple classrooms and schools are involved in the study, would be useful for pedagogists in creating new classroom materials. By evaluating student attention and engagement, they evaluate specific learning tools and subject matters. Research also suggests that the

distribution of activities has an effect on school children's attention and academic performance. Students are more likely to be attentive after engaging in physical exercise (Gapin et al., 2011; Hoza et al., 2015).

On an individual level, evaluating the attention and engagement of a student could aid in identifying subjects where the student is not challenged, struggling, or disinterested. This information would be used for creating a more personalized education plan. This would be useful under a universal design for a learning paradigm (Coyne et al., 2012; Crevecouer et al., 2014; King-Sears et al., 2015) that would allow for more individualized academic activities.

In addition to classroom activities, looking at peer and group interactions would be an important application of the approaches discussed in this dissertation. Peer teaching and group projects can be great ways for students to learn to collaborate. Providing feedback and assessments on collaboration can be difficult, and would typically require direct observation or worksheets for peer evaluation. Applying the multimodal analysis discussed in this dissertation would provide a means for extracting data from these interactions, and create video clips of both on-task and disruptive interactions.

### *8.7.3 Counseling and Therapy*

Similar to PRT interventions, counseling and therapy sessions present a dyadic scenario that would benefit from automated analysis. The techniques would be useful for gathering metrics on the treatment recipient as well as evaluating the counselor's or therapist's performance in the session. Speech and language pathology (SLP) represents an explanatory example for how the methodology could be adapted. SLP has a similar relationship to technology as autism treatments. Technological implementation research is dominated by telehealth models (Chen et al., 2016c; Ekberg et al., 2019; Keck and Doarn, 2014; Regina Molini-Avejonas et al., 2015). Other approaches have included use of ASR

systems (Lee et al., 2016), serious games (Nasiri et al., 2017), and expert tutor applications (Robles-Bykbaev et al., 2015). The approach in this dissertation is mainly applicable to the therapist-recipient relationship.

The most direct association between this dissertation and SLP therapist sessions is in extracting data to evaluate the therapist. Like PRT, the therapist's actions, in particular the auditory instructions and reinforcement techniques, could be automatically analyzed and presented. Tracking the child's responses could also be accomplished. As with child vocalizations in PRT, it is important to capture speech and non-speech vocalizations for a child in SLP therapy. This provides a means of tracking the progress between sessions that would show a visualization of the child's improvement.

Outside of therapy sessions, conversational counseling could benefit by using the automated video clip creation techniques discussed in Chapter 7. These techniques were based on instruction and response-based interactions. Using this analysis, clips could be created that isolated specific segments of the session for easier review.

### *8.7.4 Interviews*

An applicant's behavior in an interview is an important part of the hiring process for many employers. As with counseling sessions, question and response segments would be useful for hiring managers to review post-interview. Depending on the number of applicants, reviewing interview recordings could be a costly processes. Creating an interface, such as that presented in Chapter 7, would provide a simple means of reviewing specific questions during the recorded session. Additionally, body posture metrics and engagement could be gauged in a similar manner as attention detection. This would provide additional analysis of the interaction, particularly if the verbal responses were mapped to the extracted visual metrics.

Automated analysis of interview sessions has been explored in the literature. Common approaches focus on multimodal models, particularly looking at facial expression, gaze tracking, verbal language, and voice intonation (Chen et al., 2016a, 2017a; Naim et al., 2015; Rasipuram and Jayagopi, 2019). These works have primarily focused on web-based interviews. This provides the benefit of effectively having an egocentric camera perspective from the point of view of the interviewer. This affords the system the assumption of perspective, for example, looking at the camera is equivalent to looking at the interviewer. For in-person interviews, this perspective may not be able to be accommodated. In these instances, video analysis using the methodologies discussed in this dissertation would be applicable.

### *8.7.5 Police Body Cameras*

A novel and important application for the techniques researched in this dissertation is automatic video parsing of police body cameras. These cameras are worn by police officers to capture their perspective of events while on duty. Reviewing the interactions in the tapes is important for corroborating the officer's depiction of events in cases of dispute and for the community to hold authorities accountable. The data recorded in the videos is multimodal and shares similarities with the recording circumstances examined in PRT videos.

Like PRT videos, these recordings represent 'in-the-wild' conditions. It can be expected that subjects of interest will be moving in the frame, often subject to occlusion and partial depictions. Also like PRT videos, generalizations of actions are more important than recognizing the actions themselves. A subject's visual attention, his or her movements in association with other individuals, and general demeanor would be likely be more important data to extract than specific actions. By applying similar techniques to the approach used in this dissertation, automated analytics of interactions could be

collected. These analytics could be used in assessing the officer's behavior as well as identifying scenarios that would be useful in training new officers.

The expectation of having every officer recording throughout their shift creates an insurmountable amount of data for manual review. Manual review of the videos is presumably only undertaken for high profile events and disputes. Utilization of automated analysis and clip generation would provide a methodology for evaluating more data and identifying important clips for human review. Like PRT, this would benefit from keeping a human moderator as part of the system in order to gain a greater understanding of significant events that were identified by the automated system. Important events in the videos could consist of audio, video, or multimodal events. Using computer vision-based techniques would not be sufficient for creating the automated clips. Additionally, unsteady camera conditions are inevitable during periods of action. This was a concern that was discussed in Chapter 7 in regards to video summarization and clip generation. The unsteady camera conditions could cause video clip generation algorithms to misinterpret camera movement as important changes in the frame. Applying a multimodal approach using VAD and speaker separation along with normalized comparisons of body position between frames (as discussed in Chapter 7) would create more accurate clips on interactions.

Research regarding automated analysis of police body cameras has focused on activity detection (Chen et al., 2019), facial recognition (Brown and Fan, 2016) and detecting foot-chases (Aguayo et al., 2017). While these works have looked at specific aspects of detection and classification in the videos, applying the approach from this dissertation would provide more general information to aid manual review of the videos.

# REFERENCES


Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

Abushariah, M. A., Ainon, R. N., Zainuddin, R., Elshafei, M., and Khalifa, O. O. (2010). Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. In *Computer and Communication Engineering (ICCCE), 2010 International Conference on*, pages 1–6. IEEE.

Aggarwal, G. and Singh, L. (2015). Characterization between child and adult voice using machine learning algorithm. In *Computing, Communication & Automation (ICCCA), 2015 International Conference on*, pages 246–250. IEEE.

Aguayo, R., Camacho, A., Mukherjee, P., and Yang, Q. (2017). Detecting footchases from police body-worn video. *SIAM Undergraduate Research Online*, 31.

Alie, D., Mahoor, M. H., Mattson, W. I., Anderson, D. R., and Messinger, D. S. (2011). Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using markov models. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 282–287. IEEE.

Altepeter, T. S. and Breen, M. J. (1989). The Home Situations Questionnaire (HSQ) and the School Situations Questionnaire (SSQ): Normative data and an evaluation of psychometric properties. *Journal of Psychoeducational Assessment*, 7(4):312–322.

Alyamkin, S., Ardi, M., Brighton, A., Berg, A. C., Chen, Y., Cheng, H.-P., Chen, B., Fan, Z., Feng, C., Fu, B., and others (2018). 2018 low-power image recognition challenge. *arXiv preprint arXiv:1810.01732*.

Aman, M. G., Singh, N. N., Stewart, A. W., and Field, C. J. (1985). The aberrant behavior checklist: a behavior rating scale for the assessment of treatment effects. *American journal of mental deficiency*.

Andersen, S. B., Rasmussen, C. K., and Frøkjaer, E. (2017). Bringing content understanding into usability testing in complex application domains—a case study in eHealth. In *International Conference of Design, User Experience, and Usability*, pages 327–341. Springer.

Andriluka, M., Roth, S., and Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE.

Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 623–630. IEEE.

Aneeja, G. and Yegnanarayana, B. (2015). Single frequency filtering approach for discriminating speech and nonspeech. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(4):705–717.

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338.

audeering (2018). audEERING | Intelligent Audio Engineering – openSMILE. Retrieved from: https://audeering.com/technology/opensmile/.

Bachu, R., Kopparthi, S., Adapa, B., and Barkana, B. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, pages 1–7.

Bachu, R., Kopparthi, S., Adapa, B., and Barkana, B. D. (2010). Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282. Springer.

Baer, D. M., Wolf, M. M., and Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of applied behavior analysis*, 1(1):91–97.

Bagaiolo, L. F., Mari, J. d. J., Bordini, D., Ribeiro, T. C., Martone, M. C. C., Caetano, S. C., Brunoni, D., Brentani, H., and Paula, C. S. (2017). Procedures and compliance of a video modeling applied behavior analysis intervention for Brazilian parents of children with autism spectrum disorders. *Autism*, pages 603–610.

Bai, Z., Blackwell, A. F., and Coulouris, G. (2015). Using augmented reality to elicit pretend play for children with autism. *IEEE transactions on visualization and computer graphics*, 21(5):598–610.

Baker-Ericzén, M. J., Stahmer, A. C., and Burns, A. (2007). Child demographics associated with outcomes in a community-based pivotal response training program. *Journal of positive behavior interventions*, 9(1):52–60.

Baxter, R. H., Leach, M. J., Mukherjee, S. S., and Robertson, N. M. (2015). An adaptive motion model for person tracking with instantaneous head-pose features. *IEEE Signal Processing Letters*, 22(5):578–582.

Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., and Murino, V. (2013). Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127.

Beals, C. (2017). Augmented Reality Sandbox will Blow Your Mind! Retrieved from: https://www.youtube.com/watch?v=ba4uvkastpc&feature=youtu.be.

Benko, H., Wilson, A. D., and Zannier, F. (2014). Dyadic projected spatial augmented reality. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 645–655. ACM.

Bertero, D., Fung, P., Li, X., Wu, L., LIU, Z., Hussain, B., Chong, W., Lau, K., Yue, P., Zhang, W., and others (2016). Deep Learning of Audio and Language Features for Humor Prediction. In *LREC*.

Billeci, L., Tonacci, A., Tartarisco, G., Narzisi, A., Di Palma, S., Corda, D., Baldus, G., Cruciani, F., Anzalone, S. M., Calderoni, S., and others (2016). An integrated approach for the monitoring of brain and autonomic response of children with autism spectrum disorders during treatment by wearable technologies. *Frontiers in neuroscience*, 10:276.

Billinghurst, M. and Kato, H. (2002). Collaborative augmented reality. *Communications of the ACM*, 45(7):64–70.

Bölte, S., Bartl-Pokorny, K., Jonsson, U., Berggren, S., Zhang, D., Kostrzewa, E., Falck-Ytter, T., Einspieler, C., Pokorny, F., Jones, E., and others (2016). How can clinicians detect and treat autism early? Methodological trends of technology use in research. *Acta Paediatrica*, 105(2):137–144.

Boersma, P. and Weenink, D. (2018). Praat: doing Phonetics by Computer.

Bootkrajang, J. and Kabán, A. (2014). Learning kernel logistic regression in the presence of class label noise. *Pattern Recognition*, 47(11):3641–3655.

Boril, H., Zhang, Q., Ziaei, A., Hansen, J. H., Xu, D., Gilkerson, J., Richards, J. A., Zhang, Y., Xu, X., Mao, H., and others (2014). Automatic assessment of language background in toddlers through phonotactic and pitch pattern modeling of short vocalizations. In *WOCCI*, pages 39–43.

Bosch, N., Chen, H., D'Mello, S., Baker, R., and Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 267–274. ACM.

Boukadida, H., Berrani, S.-A., and Gros, P. (2017). Automatically creating adaptive video summaries using constraint satisfaction programming: Application to sport content. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):920–934.

Brookman-Frazee, L. and Koegel, R. (2004). Using parent/clinician partnerships in parent education programs for children with autism. *Journal of positive behavior interventions*, 6(4):195–213.

Brown, L. M. and Fan, Q. (2016). Enhanced face detection using body part detections for wearable cameras. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 715–720. IEEE.

Brutzer, S., Høferlin, B., and Heidemann, G. (2011). Evaluation of background subtraction techniques for video surveillance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1937–1944. IEEE.

Bu, S., Zhao, Y., Hwang, M.-Y., and Sun, S. (2018). A Probability Weighted Beamformer for Noise Robust ASR. In *Interspeech*, pages 3048–3052.

Campos, V., Jou, B., and Giro-i Nieto, X. (2017). From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing*, 65:15–22.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields. In *CVPR*.

Casale, P., Pujol, O., and Radeva, P. (2011). Human activity recognition from accelerometer data using a wearable device. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 289–296. Springer.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and McOwan, P. W. (2012). Detecting engagement in HRI: An exploration of social and task-based context. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 421–428. IEEE.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960–4964. IEEE.

Charles, J., Pfister, T., Magee, D., Hogg, D., and Zisserman, A. (2016). Personalizing human video pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3063–3072.

Chen, C. Y. and Grauman, K. (2017). Efficient Activity Detection in Untrimmed Video with Max-Subgraph Search. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):908–921.

Chen, H., Li, H., Song, A., Haberland, M., Akar, O., Dhillon, A., Zhou, T., Bertozzi, A. L., and Brantingham, P. J. (2019). Semi-Supervised First-Person Activity Recognition in Body-Worn Video. *arXiv preprint arXiv:1904.09062*.

Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., and Yoon, S.-Y. (2016a). Automated scoring of interview videos using Doc2vec multimodal feature extraction paradigm. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 161–168. ACM.

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., and Hoque, M. E. (2017a). Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE.

Chen, X., Liu, X., Wang, Y., Gales, M. J., and Woodland, P. C. (2016b). Efficient training and evaluation of recurrent neural network language models for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2146–2157.

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., and others (2016c). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language*, 37:98–128.

Chen, Z., Luo, Y., and Mesgarani, N. (2017b). Deep attractor network for single-microphone speaker separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250. IEEE.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Chiang, H.-H., Wu, S.-J., Perng, J.-W., Wu, B.-F., and Lee, T.-T. (2010). The human-in-the-loop design approach to the longitudinal automation system for an intelligent vehicle. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(4):708–720.

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., and Darrell, T. (2016). Best practices for fine-tuning visual classifiers to new domains. In *European conference on computer vision*, pages 435–442. Springer.

Coe, D., Matson, J., Fee, V., Manikam, R., and Linarello, C. (1990). Training nonverbal and verbal play skills to mentally retarded and autistic children. *Journal of autism and developmental disorders*, 20(2):177–187.

Conati, C., Chabbal, R., and Maclaren, H. (2003). A study on using biometric sensors for monitoring user emotions in educational games. In *Workshop on Assessing and Adapting to User Attitudes and Affect: Why, When and How*. Citeseer.

Considine, B. (2011). Incidental Teaching Retrieved from: https://www.youtube.com/watch?v=vwoayir7vsk.

Constantino, J. N. (2013). *Social responsiveness scale*. Springer.

Coolican, J., Smith, I. M., and Bryson, S. E. (2010). Brief parent training in pivotal response treatment for preschoolers with autism. *Journal of Child Psychology and Psychiatry*, 51(12):1321–1330.

Coronato, A., De Pietro, G., and Paragliola, G. (2014). A situation-aware system for the detection of motion disorders of patients with Autism Spectrum Disorders. *Expert Systems with Applications*, 41(17):7868–7877.

Coyne, P., Pisha, B., Dalton, B., Zeph, L. A., and Smith, N. C. (2012). Literacy by Design A Universal Design for Learning Approach for Students With Significant Intellectual Disabilities. *Remedial and Special Education*, 33(3):162–172.

Crevecouer, Y., Sorenson, S. E., Mayorga, V., and Gonzalez, A. P. (2014). Universal design for learning in K-12 educational settings: A review of group comparison and single-subject intervention studies. *The Journal of Special Education Apprenticeship*, 3(2):1–23.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4.

Debenedictis, E., Lu, Y.-H., Kadin, A., Berg, A., Conte, T., Garg, R., Gingade, G., Hoang, B., Huang, Y., Li, B., and others (2016). Rebooting Computing and Low-Power Image Recognition Challenge. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Demir, M. and Isil Bozma, H. (2015). Video summarization via segments summary graphs. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–25.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Deng, Z., Vahdat, A., Hu, H., and Mori, G. (2016). 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*, pages 4772–4781.

Deng, Z., Zhai, M., Chen, L., Liu, Y., Muralidharan, S., Roshtkhari, M. J., and Mori, G. (2015). Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191*.

DeVries, T. and Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Dierker, A., Pitsch, K., and Hermann, T. (2011). An augmented-reality-based scenario for the collaborative construction of an interactive museum.

D'Mello, S. K. (2016). Giving eyesight to the blind: Towards attention-aware AIED. *International Journal of Artificial Intelligence in Education*, 26(2):645–659.

D'mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, 18(1-2):45–80.

D'mello, S. K. and Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187.

Dorfmüller, K. (1999). Robust tracking for augmented reality using retroreflective markers. *Computers & Graphics*, 23(6):795–800.

Drugman, T., Stylianou, Y., Kida, Y., and Akamine, M. (2016). Voice activity detection: Merging source and filter-based information. *IEEE Signal Processing Letters*, 23(2):252–256.

Dubagunta, S. P., Kabil, S. H., and Doss, M. M. (2019). Improving Children Speech Recognition through Feature Learning from Raw Speech Signal.

Dudy, S., Bedrick, S., Asgari, M., and Kain, A. (2017). Automatic analysis of pronunciations for children with speech sound disorders. *Computer Speech & Language*.

Duffner, S. and Garcia, C. (2016). Visual focus of attention estimation with unsupervised incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(12):2264–2272.

Einfalt, M., Zecha, D., and Lienhart, R. (2018). Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 446–455. IEEE.

Ekberg, S., Danby, S., Theobald, M., Fisher, B., and Wyeth, P. (2019). Using physical objects with young children in 'face-to-face'and telehealth speech and language therapy. *Disability and rehabilitation*, 41(14):1664–1675.

El-Baz, A., Casanova, M. F., Gimel'farb, G., Mott, M., and Switala, A. E. (2007). Autism diagnostics by 3d texture analysis of cerebral white matter gyrifications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 882–890. Springer.

Elnakib, A., Casanova, M. F., Gimel'farb, G., Switala, A. E., and El-Baz, A. (2011). Autism diagnostics by centerline-based shape analysis of the corpus callosum. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1843–1846. IEEE.

Enqing, D., Guizhong, L., Yatong, Z., and Xiaodi, Z. (2002). Applying support vector machines to voice activity detection. In *Signal Processing, 2002 6th International Conference on*, volume 2, pages 1124–1127. IEEE.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Escobedo, L., Tentori, M., Quintana, E., Favela, J., and Garcia-Rosas, D. (2014). Using augmented reality to help children with autism stay focused. *IEEE Pervasive Computing*, 13(1):38–46.

Estes, A., Vismara, L., Mercado, C., Fitzpatrick, A., Elder, L., Greenson, J., Lord, C., Munson, J., Winter, J., Young, G., and others (2014). The impact of parent-delivered intervention on parents of very young children with autism. *Journal of autism and developmental disorders*, 44(2):353–365.

Eyben, F., Wöllmer, M., and Schuller, B. (2009). OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*, pages 1–6. IEEE.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79.

Feng, L., Wiltsche, C., Humphrey, L., and Topcu, U. (2016). Synthesis of human-in-the-loop control protocols for autonomous systems. *IEEE Transactions on Automation Science and Engineering*, 13(2):450–462.

Fenson, L. and others (2007). *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.

Fletcher, R. R., Ming-Zher Poh, R., and Eydgahi, H. (2010). Wearable sensors: opportunities and challenges for low-cost health care.

Foster, M. E., Gaschler, A., and Giuliani, M. (2013). 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*, pages 255–262. ACM.

Fragkiadaki, K., Levine, S., Felsen, P., and Malik, J. (2015). Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354.

Frank, M. C., Simmons, K., Yurovsky, D., and Pusiol, G. (2013). Developmental and postural changes in children's visual access to faces. In *CogSci*.

Gabbay, A., Ephrat, A., Halperin, T., and Peleg, S. (2018). Seeing through noise: Visually driven speaker separation and enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3051–3055. IEEE.

Gapin, J. I., Labban, J. D., and Etnier, J. L. (2011). The effects of physical activity on attention deficit hyperactivity disorder symptoms: the evidence. *Preventive Medicine*, 52:S70–S74.

Garcia, K. D., Carvalho, T., Mendes-Moreira, J., Cardoso, J. M., and de Carvalho, A. C. (2018). A Preliminary Study on Hyperparameter Configuration for Human Activity Recognition. *arXiv preprint arXiv:1810.10956*.

Gauen, K., Rangan, R., Mohan, A., Lu, Y.-H., Liu, W., and Berg, A. C. (2017). Low-power image recognition challenge. In *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 99–104. IEEE.

Ge, W. and Yu, Y. (2017). Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1086–1095.

Geifman, Y., Uziel, G., and El-Yaniv, R. (2018). Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers.

Gengoux, G. W., Berquist, K. L., Salzman, E., Schapp, S., Phillips, J. M., Frazier, T. W., Minjarez, M. B., and Hardan, A. Y. (2015). Pivotal response treatment parent training for autism: Findings from a 3-month follow-up evaluation. *Journal of autism and developmental disorders*, 45(9):2889–2898.

Giannakopoulos, T. (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one*, 10(12).

Gillesen, J. C., Barakova, E., Huskens, B. E., and Feijs, L. M. (2011). From training to robot behavior: Towards custom scenarios for robotics in training programs for ASD. In *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, pages 1–7. IEEE.

Gillett, J. N. and LeBlanc, L. A. (2007). Parent-implemented natural language paradigm to increase language and play in children with autism. *Research in Autism Spectrum Disorders*, 1(3):247–255.

Golik, P., Tüske, Z., Schlüter, R., and Ney, H. (2015). Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In *Sixteenth annual conference of the international speech communication association*.

Gordon, M., Henderson, R., Holmes, J. H., Wolters, M. K., Bennett, I. M., and Group, S. S. i. P. I. R. w. I. T. (2015). Participatory design of ehealth solutions for women from vulnerable populations with perinatal depression. *Journal of the American Medical Informatics Association*, 23(1):105–109.

Gosztolya, G. (2016). Detecting laughter and filler events by time series smoothing with genetic algorithms. *International Conference on Speech and Computer*, pages 232–239.

Gosztolya, G., Grósz, T., Busa-Fekete, R., and Tóth, L. (2016a). Determining native language and deception using phonetic features and classifier combination. *Interspeech*.

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., and Kálmán, J. (2016b). Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. *Computer Speech & Language*.

Graf, S., Herbig, T., Buck, M., and Schmidt, G. (2015). Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):91.

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., and Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 42–49. ACM.

Graves, A., Mohamed, A., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutierrez-Santos, S., Wiedmann, M., and Rummel, N. (2016). Affecting off-task behaviour: how affect-aware feedback can improve student learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 104–113. ACM.

Griffiths, R. (1996). *The Griffiths Mental Development Scales: From Birth to 2 Years: Manual*. The Test Agency.

Górriz, J. M., Ramírez, J., Lang, E. W., and Puntonet, C. G. (2006). Hard c-means clustering for voice activity detection. *Speech communication*, 48(12):1638–1649.

Gruyer, D., Magnier, V., Hamdi, K., Claussmann, L., Orfila, O., and Rakotonirainy, A. (2017). Perception, information processing and modeling: Critical stages for autonomous driving applications. *Annual Reviews in Control*, 44:323–341.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

Gutierrez Jr, A., Hale, M. N., O'Brien, H. A., Fischer, A. J., Durocher, J. S., and Alessandri, M. (2009). Evaluating the effectiveness of two commonly used discrete trial procedures for teaching receptive discrimination to young children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 3(3):630–638.

Guy, W. (1976). ECDEU assessment manual for psychopharmacology. *US Department of Health, and Welfare*, pages 534–537.

Hardan, A. Y., Gengoux, G. W., Berquist, K. L., Libove, R. A., Ardel, C. M., Phillips, J., Frazier, T. W., and Minjarez, M. B. (2015). A randomized controlled trial of Pivotal Response Treatment Group for parents of children with autism. *Journal of Child Psychology and Psychiatry*, 56(8):884–892.

Harper, C. B., Symon, J. B., and Frea, W. D. (2008). Recess is time-in: Using peers to improve social skills of children with autism. *Journal of autism and developmental disorders*, 38(5):815–826.

Harwath, D. and Glass, J. R. (2017). Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*.

Harwath, D., Torralba, A., and Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Hashemi, J., Spina, T. V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., and Sapiro, G. (2012). A computer vision approach for the assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE.

Heath, C. D., Heath, T., McDaniel, T., Venkateswara, H., and Panchanathan, S. (2019a). Using Participatory Design to Create a User Interface for Analyzing Pivotal Response Treatment Video Probes. In *International Conference on Smart Multimedia [Accepted]*.

Heath, C. D., McDaniel, T., Venkateswara, H., and Panchanathan, S. (2019b). Improving Communication Skills of Children with Autism through Support of Applied Behavioral Analysis Treatments using Multimedia Computing: A Survey. *Universal Access in the Information Society [Accepted]*.

Heath, C. D., McDaniel, T., Venkateswara, H., and Panchanathan, S. (2019c). Parent and Child Voice Activity Detection in Pivotal Response Treatment Video Probes. *Human Computer Interaction International*.

Heath, C. D., Venkateswara, H., McDaniel, T., and Panchanathan, S. (2018). Detecting Attention in Pivotal Response Treatment Video Probes. In *International Conference on Smart Multimedia*.

Heath, C. D., Venkateswara, H., McDaniel, T., and Panchanathan, S. (2019d). Using Multimodal Data for Automated Fidelity Evaluation in Pivotal Response Treatment Videos. In *Global Signal and Information Processing [Accepted]*.

Heath, C. D., Venkateswara, H., and Panchanathan, S. (2019e). Are You Paying Attention? Classifying Attention in Pivotal Response Treatment Videos. *CVPR - Workshop on LPIRC*.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and others (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hönig, F., Stemmer, G., Hacker, C., and Brugnara, F. (2005). Revising perceptual linear prediction (PLP). In *Ninth European Conference on Speech Communication and Technology*.

Hoai, M. and Zisserman, A. (2014). Talking heads: Detecting humans and recognizing their interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 875–882.

Holroyd, J. (1974). The Questionnaire on Resources and Stress: An instrument to measure family response to a handicapped family member. *Journal of community psychology*, 2(1):92–94.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., and Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202.

Hoshen, Y., Weiss, R. J., and Wilson, K. W. (2015). Speech acoustic modeling from raw multichannel waveforms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4624–4628. IEEE.

Hoza, B., Smith, A. L., Shoulberg, E. K., Linnea, K. S., Dorsch, T. E., Blazo, J. A., Alerding, C. M., and McCabe, G. P. (2015). A randomized trial examining the effects of aerobic physical activity on attention-deficit/hyperactivity disorder symptoms in young children. *Journal of abnormal child psychology*, 43(4):655–667.

Hughes, T. and Mierle, K. (2013). Recurrent neural networks for voice activity detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7378–7382. IEEE.

Hunter, E. J. (2009). A comparison of a child's fundamental frequencies in structured elicited vocalizations versus unstructured natural vocalizations: A case study. *International journal of pediatric otorhinolaryngology*, 73(4):561–571.

Isaacson, R. and Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning. *Journal of the Scholarship of Teaching and Learning*, pages 39–55.

Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*.

Ito, M. and Donaldson, R. (1971). Zero-crossing measurements for analysis and recognition of speech sounds. *IEEE Transactions on Audio and Electroacoustics*, 19(3):235–242.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Jain, A., Zamir, A. R., Savarese, S., and Saxena, A. (2016). Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317.

Jaitly, N., Nguyen, P., Senior, A., and Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Jazouli, M., Elhoufi, S., Majda, A., Zarghili, A., and Aalouane, R. (2016). Stereotypical Motor Movement Recognition Using Microsoft Kinect with Artificial Neural Network. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 10(7):1270–1274.

Jiang, M. and Zhao, Q. (2017). Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3267–3276.

Jindal, I., Nokleby, M., and Chen, X. (2016). Learning deep networks from noisy labels with dropout regularization. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 967–972. IEEE.

Jo, Q.-H., Chang, J.-H., Shin, J., and Kim, N. (2009). Statistical model-based voice activity detection using support vector machine. *IET Signal Processing*, 3(3):205–210.

Johnson, N., Frenn, M., Feetham, S., and Simpson, P. (2011). Autism spectrum disorder: Parenting stress, family functioning and health-related quality of life. *Families, systems, & health*, 29(3):232.

Johnston, C. and Mash, E. J. (1989). A measure of parenting satisfaction and efficacy. *Journal of clinical child psychology*, 18(2):167–175.

Jones, E. A. and Feeley, K. M. (2009). Parent implemented joint attention intervention for preschoolers with autism. *Play and Social Skills for Children with Autism Spectrum Disorder*.

Jong, M. d., Mavridis, P., Aroyo, L., Bozzon, A., Vos, J. d., Oomen, J., Dimitrova, A., and Badenoch, A. (2018). A Human in the Loop Approach to Capture Bias and Support Media Scientists in News Video Analysis. *Joint Proceedings SAD 2018 and CrowdBias 2018*, 2276:32–40.

Ju, Y., Li, L., Jiao, L., Ren, Z., Hou, B., and Yang, S. (2018). Modified Diversity of Class Probability Estimation Co-training for Hyperspectral Image Classification. *arXiv preprint arXiv:1809.01436*.

Kane, M., Connell, J. E., and Pellecchia, M. (2010). A quantitative analysis of language interventions for children with autism. *The Behavior Analyst Today*, 11(2):128.

Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., and Berry, K. (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of consulting and clinical psychology*, 83(3):554.

Kato, N., Li, T., Nishino, K., and Uchida, Y. (2018). Improving Multi-Person Pose Estimation using Label Correction. *arXiv preprint arXiv:1811.03331*.

Kazdin, A. E. and Whitley, M. K. (2003). Treatment of parental stress to enhance therapeutic change among children referred for aggressive and antisocial behavior. *Journal of consulting and clinical psychology*, 71(3):504.

Keck, C. S. and Doarn, C. R. (2014). Telehealth technology applications in speech-language pathology. *Telemedicine and e-Health*, 20(7):653–659.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Khan, N. A., Sawand, M. A., Qadeer, M., Owais, A., Junaid, S., and Shahnawaz, P. (2017). Autism Detection using Computer Vision. *International Journal of Computer Science and Network Security (IJCSNS)*, 17(4):256.

Kientz, J. A., Hayes, G. R., Westeyn, T. L., Starner, T., and Abowd, G. D. (2007). Pervasive computing and autism: Assisting caregivers of children with special needs. *IEEE Pervasive Computing*, 6(1):28–35.

Kildea, J., Battista, J., Cabral, B., Hendren, L., Herrera, D., Hijal, T., and Joseph, A. (2019). Design and Development of a Person-Centered Patient Portal Using Participatory Stakeholder Co-Design. *Journal of medical Internet research*, 21(2):e11371.

Kim, J. and Hahn, M. (2018). Voice Activity Detection Using an Adaptive Context Attention Model. *IEEE Signal Processing Letters*, 25(8):1181.

King-Sears, M. E., Johnson, T. M., Berkeley, S., Weiss, M. P., Peters-Burton, E. E., Evmenova, A. S., Menditto, A., and Hursh, J. C. (2015). An exploratory study of universal design for teaching chemistry to students with and without disabilities. *Learning Disability Quarterly*, 38(2):84–96.

Kitsantas, A. and Kavussanu, M. (2011). Acquisition of sport knowledge and skill. *Handbook of self-regulation of learning and performance*, pages 217–233.

Koegel, L. K., Camarata, S. M., Valdez-Menchaca, M., and Koegel, R. L. (1997). Setting generalization of question-asking by children with autism. *American Journal on Mental Retardation*, 102(4):346–357.

Koegel, L. K., Carter, C. M., and Koegel, R. L. (2003). Teaching children with autism self-initiations as a pivotal response. *Topics in language disorders*, 23(2):134–145.

Koegel, L. K., Koegel, R. L., Green-Hopkins, I., and Barnes, C. C. (2010). Brief report: Question-asking and collateral language acquisition in children with autism. *Journal of Autism and Developmental Disorders*, 40(4):509–515.

Koegel, L. K., Koegel, R. L., Harrower, J. K., and Carter, C. M. (1999a). Pivotal response intervention I: Overview of approach. *Journal of the Association for Persons with Severe Handicaps*, 24(3):174–185.

Koegel, L. K., Koegel, R. L., Shoshan, Y., and McNerney, E. (1999b). Pivotal response intervention II: Preliminary long-term outcome data. *Journal of the Association for Persons with Severe Handicaps*, 24(3):186–198.

Koegel, L. K., Singh, A. K., Koegel, R. L., Hollingsworth, J. R., and Bradshaw, J. (2014a). Assessing and improving early social engagement in infants. *Journal of positive behavior interventions*, 16(2):69–80.

Koegel, R. L. (1988). *How To Teach Pivotal Behaviors to Children with Autism: A Training Manual.*

Koegel, R. L., Bimbela, A., and Schreibman, L. (1996). Collateral effects of parent training on family interactions. *Journal of autism and developmental disorders*, 26(3):347–359.

Koegel, R. L., Bradshaw, J. L., Ashbaugh, K., and Koegel, L. K. (2014b). Improving question-asking initiations in young children with autism using pivotal response treatment. *Journal of autism and developmental disorders*, 44(4):816–827.

Koegel, R. L., Camarata, S., Koegel, L. K., Ben-Tall, A., and Smith, A. E. (1998). Increasing speech intelligibility in children with autism. *Journal of autism and developmental disorders*, 28(3):241–251.

Koegel, R. L., Koegel, L. K., and Surratt, A. (1992). Language intervention and disruptive behavior in preschool children with autism. *Journal of autism and developmental disorders*, 22(2):141–153.

Koegel, R. L., O'Dell, M., and Dunlap, G. (1988). Producing speech use in nonverbal autistic children by reinforcing attempts. *Journal of autism and developmental disorders*, 18(4):525–538.

Koegel, R. L., O'dell, M. C., and Koegel, L. K. (1987). A natural language teaching paradigm for nonverbal autistic children. *Journal of autism and developmental disorders*, 17(2):187–200.

Koegel, R. L., Symon, J. B., and Kern Koegel, L. (2002). Parent education for families of children with autism living in geographically distant areas. *Journal of Positive Behavior Interventions*, 4(2):88–103.

Koegel, R. L., Vernon, T. W., and Koegel, L. K. (2009). Improving social initiations in young children with autism using reinforcers with embedded social interactions. *Journal of autism and developmental disorders*, 39(9):1240–1251.

Koh, Y. J. and Kim, C.-S. (2017). Primary Object Segmentation in Videos Based on Region Augmentation and Reduction. In *CVPR*, volume 1, page 7.

Kratzert, F. (2017). Finetuning AlexNet with TensorFlow. Retrieved from: https://kratzert.github.io/2017/02/24/finetuning-alexnet-with-tensorflow.html.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kuhl, J. (2000). A functional-design approach to motivation and self-regulation: The dynamics of personality systems interactions. In *Handbook of self-regulation*, pages 111–169. Elsevier.

Kumar, A., Kim, J., Lyndon, D., Fulham, M., and Feng, D. (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics*, 21(1):31–40.

Kumar, M., Bone, D., McWilliams, K., Williams, S., Lyon, T. D., and Narayanan, S. (2017). Multi-scale Context Adaptation for Improving Child Automatic Speech Recognition in Child-Adult Spoken Interactions. *Proc. Interspeech 2017*, pages 2730–2734.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Laski, K. E., Charlop, M. H., and Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, 21(4):391–400.

Lawton, K. and Kasari, C. (2012). Teacher-implemented joint attention intervention: Pilot randomized controlled study for preschoolers with autism. *Journal of consulting and clinical psychology*, 80(4):687.

Le, D., Aldeneh, Z., and Provost, E. M. (2017). Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network. In *INTERSPEECH*, pages 1108–1112.

Leaf, J. B., Leaf, R., Alcalay, A., Leaf, J. A., Ravid, D., Dale, S., Kassardjian, A., Tsuji, K., Taubman, M., McEachin, J., and others (2015). Utility of formal preference assessments for individuals diagnosed with autism spectrum disorder. *Education and Training in Autism and Developmental Disabilities*, 50(2):199.

Leaf, J. B., Leaf, R., McEachin, J., Taubman, M., Ala'i-Rosales, S., Ross, R. K., Smith, T., and Weiss, M. J. (2016). Applied behavior analysis is a science and, therefore, progressive. *Journal of autism and developmental disorders*, 46(2):720–731.

Lecavalier, L., Smith, T., Johnson, C., Bearss, K., Swiezy, N., Aman, M. G., Sukhodolsky, D. G., Deng, Y., Dziura, J., and Scahill, L. (2017). Moderators of parent training for disruptive behaviors in young children with autism spectrum disorder. *Journal of abnormal child psychology*, 45(6):1235–1245.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Lee, T., Liu, Y., Huang, P.-W., Chien, J.-T., Lam, W. K., Yeung, Y. T., Law, T. K., Lee, K. Y., Kong, A. P.-H., and Law, S.-P. (2016). Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479. IEEE.

Lewis, C. and Rieman, J. (1993). Task-centered user interface design. *A practical introduction*.

Li, F., Zhang, G., Wang, W., Xu, R., Schnell, T., Wen, J., McKenzie, F., and Li, J. (2017). Deep models for engagement assessment with scarce label information. *IEEE Transactions on Human-Machine Systems*, 47(4):598–605.

Li, G. (2017). Human-in-the-loop data integration. *Proceedings of the VLDB Endowment*, 10(12):2006–2017.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Li, L., Xu, Q., and Tan, Y. K. (2012). Attention-based addressee selection for service and social robots to interact with multiple persons. In *Proceedings of the Workshop at SIGGRAPH Asia*, pages 131–136. ACM.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., and Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1611–1615.

Lim, C. and Park, T. (2011). Exploring the educational use of an augmented reality books. In *Proceedings of the Annual Convention of the Association for Educational Communications and Technology*, pages 172–182.

Liu, R., Salisbury, J. P., Vahabzadeh, A., and Sahin, N. T. (2017). Feasibility of an autism-focused augmented reality smartglasses system for social communication and behavioral coaching. *Frontiers in pediatrics*, 5:145.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3):205–223.

Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685.

Lovaas, O. I. and Smith, T. (2003). Early and intensive behavioral intervention in autism. *Evidence-based Psychotherapies for Children and Adolescents*, pages 328–340.

Lu, Y.-H. (2019). Low-power image recognition. *Nature Machine Intelligence*, 1(4):199.

Lu, Y.-H., Kadin, A. M., Berg, A. C., Conte, T. M., DeBenedictis, E. P., Garg, R., Gingade, G., Hoang, B., Huang, Y., Li, B., and others (2015). Rebooting computing and low-power image recognition challenge. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 927–932. IEEE.

Ma, S., Sigal, L., and Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1942–1950.

Machalicek, W., O'Reilly, M. F., Rispoli, M., Davis, T., Lang, R., Franco, J. H., and Chan, J. M. (2010). Training teachers to assess the challenging behaviors of students with autism using video tele-conferencing. *Education and Training in Autism and Developmental Disabilities*, pages 203–215.

Mahasseni, B., Lam, M., and Todorovic, S. (2017). Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211.

Marchi, E., Schuller, B., Baron-Cohen, S., Golan, O., Bölte, S., Arora, P., and Häb-Umbach, R. (2015). 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*, pages 115–119.

Marcu, G., Dey, A. K., and Kiesler, S. (2012). Parent-driven use of wearable cameras for autism support: a field study with families. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 401–410. ACM.

Martin, K. B., Hammal, Z., Ren, G., Cohn, J. F., Cassell, J., Ogihara, M., Britton, J. C., Gutierrez, A., and Messinger, D. S. (2018). Objective measurement of head movement differences in children with and without autism spectrum disorder. *Molecular autism*, 9(1):14.

Matson, J. and LoVullo, S. (2008). A review of behavioral treatments for self-injurious behaviors of persons with autism spectrum disorders. *Behavior Modification*, 32(1):61–76.

McClelland, A., Jenson, W. R., Clark, E., Davis, J., Director, G., and Hood, J. (2016). Comparisons of Pivotal Response Treatment (PRT) and Discrete Trial Training (DTT), PhD Dissertation, University of Utah, Salt Lake City, utah.

McLoughlin, I. V. (2014). The use of low-frequency ultrasound for voice activity detection. In *Fifteenth Annual Conference of the International Speech Communication Association*.

McMahon, D., Cihak, D. F., and Wright, R. (2015). Augmented reality as a navigation tool to employment opportunities for postsecondary education students with intellectual disabilities and autism. *Journal of Research on Technology in Education*, 47(3):157–172.

Mehner, W., Boltes, M., Mathias, M., and Leibe, B. (2015). Robust marker-based tracking for measuring crowd dynamics. In *International Conference on Computer Vision Systems*, pages 445–455. Springer.

Mei, S., Guan, G., Wang, Z., Wan, S., He, M., and Feng, D. D. (2015). Video summarization via minimum sparse reconstruction. *Pattern Recognition*, 48(2):522–533.

Mendels, G., Levitan, S. I., Lee, K.-Z., and Hirschberg, J. (2017). Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. In *Proc. Interspeech 2017*, pages 1472–1476.

Meng, J., Wang, H., Yuan, J., and Tan, Y.-P. (2016). From keyframes to key objects: Video summarization by representative object proposal selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1039–1048.

Miao, Y., Gowayyed, M., and Metze, F. (2015). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 167–174. IEEE.

Mitchell, K. M., Holtz, B. E., and McCarroll, A. (2018). Patient-centered methods for designing and developing health information communication technologies: a systematic review. *Telemedicine and e-Health*.

Moencks, M., De Silva, V., Roche, J., and Kondoz, A. (2019). Adaptive Feature Processing for Robust Human Activity Recognition on a Novel Multi-Modal Dataset. *arXiv preprint arXiv:1901.02858*.

Mohammadzaheri, F., Koegel, L. K., Rezaee, M., and Rafiee, S. M. (2014). A randomized clinical trial comparison between pivotal response treatment (PRT) and structured applied behavior analysis (ABA) intervention for children with autism. *Journal of autism and developmental disorders*, 44(11):2769–2777.

Mohammadzaheri, F., Koegel, L. K., Rezaei, M., and Bakhshi, E. (2015). 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. *Journal of autism and developmental disorders*, 45(9):2899–2907.

Moore, R., Caines, A., Graham, C., and Buttery, P. (2016). Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 782–793.

Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M. (1997). Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing*, 5(5):407–424.

Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE.

Mullen, E. M. and others (1995). *Mullen scales of early learning*. AGS Circle Pines, MN.

Naim, I., Tanveer, M. I., Gildea, D., and Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–6. IEEE.

Narayanan, A. and Wang, D. (2014). Joint noise adaptive training for robust automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2504–2508. IEEE.

Nasiri, N., Shirmohammadi, S., and Rashed, A. (2017). A serious game for children with speech disorders and hearing problems. In *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–7. IEEE.

Nazneen, N., Oberleitner, R., and Reischl, U. (2017). Asynchronous Telemedicine System for the Remote Diagnosis of ASD. In *Telemedicine*, pages 1–14. SM Journals.

Nazneen, N., Rozga, A., Smith, C. J., Oberleitner, R., Abowd, G. D., and Arriaga, R. I. (2015). A novel system for supporting autism diagnosis using home videos: iterative development and evaluation of system design. *JMIR mHealth and uHealth*, 3(2):e68.

Nefdt, N., Koegel, R., Singer, G., and Gerber, M. (2010). The use of a self-directed learning program to provide introductory training in pivotal response treatment to parents of children with autism. *Journal of Positive Behavior Interventions*, 12(1):23–32.

Neumann, L., Zisserman, A., and Vedaldi, A. (2018). Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection. *NIPS Workshop on Machine Learning for Intelligent Transportation Systems.*

Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J. (2005). Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):296–305.

Oberleitner, R., Reischl, U., Gazieva, K. G., Nazneen, N., Suri, J. S., and Smith, C. J. (2017). Behavior Imaging®: Resolving assessment challenges for autism spectrum disorder in pharmaceutical trials. In *Autism Imaging and Devices*, pages 371–386. CRC Press.

Orr, E. and Geva, R. (2015). Symbolic play and language development. *Infant behavior and development*, 38:147–161.

O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE potentials*, 7(1):29–32.

Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648.

Palaz, D., Collobert, R., and Doss, M. M. (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*.

Palaz, D., Collobert, R., and others (2015a). Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap.

Palaz, D., Doss, M. M., and Collobert, R. (2015b). Convolutional neural networks-based continuous speech recognition using raw speech signal. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4295–4299. IEEE.

Parthasarathy, S. and Busso, C. (2017). Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In *INTERSPEECH*, pages 1103–1107.

Passricha, V. and Aggarwal, R. K. (2018). Convolutional Neural Networks for Raw Speech Recognition. In *From Natural to Artificial Intelligence-Algorithms and Applications*. IntechOpen.

Patron-Perez, A., Marszalek, M., Zisserman, A., and Reid, I. D. (2010). High Five: Recognising human interactions in TV shows. In *BMVC*, volume 1, page 2. Citeseer.

Pawar, R., Albin, A., Gupta, U., Rao, H., Carberry, C., Hamo, A., Jones, R. M., Lord, C., and Clements, M. A. (2017). Automatic analysis of LENA recordings for language assessment in children aged five to fourteen years with application to individuals with autism. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 245–248. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pejsa, T., Kantor, J., Benko, H., Ofek, E., and Wilson, A. (2016). Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1716–1725. ACM.

Peternel, L., Petrič, T., and Babič, J. (2015). Human-in-the-loop approach for teaching robot assembly tasks using impedance control interface. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1497–1502. IEEE.

Peternel, L., Petrič, T., and Babič, J. (2018). Robotic assembly solution by human-in-the-loop teaching method based on real-time stiffness modulation. *Autonomous Robots*, 42(1):1–17.

Pi, J., Gu, Y., Hu, K., Cheng, X., Zhan, Y., and Wang, Y. (2016). Real-time scale-adaptive correlation filters tracker with depth information to handle occlusion. *Journal of Electronic Imaging*, 25(4):043022.

Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3575–3584.

Pierce, K. and Schreibman, L. (1995). Increasing complex social behaviors in children with autism: Effects of Peer-implemented pivotal response training. *Journal of applied behavior analysis*, 28(3):285–295.

Pierce, K. and Schreibman, L. (1997). Multiple peer use of pivotal response training to increase social behaviors of classmates with autism: Results from trained and untrained peers. *Journal of applied behavior analysis*, 30(1):157–160.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation*, pages 451–502. Elsevier.

Pishchulin, L., Andriluka, M., Gehler, P., and Schiele, B. (2013). Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

Potamianos, A. and Narayanan, S. (1998). Spoken dialog systems for children. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 197–200. IEEE.

Presti, L., Sclaroff, S., and Rozga, A. (2013). Joint alignment and modeling of correlated behavior streams. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 730–737.

Pusiol, G., Soriano, L., Frank, M. C., and Fei-Fei, L. (2014). Discovering the signatures of joint attention in child-caregiver interaction. In *Proceedings of the Cognitive Science Society*, volume 36.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raca, M., Kidzinski, L., and Dillenbourg, P. (2015). Translating head motion into attention-towards processing of student's body-language. In *Proceedings of the 8th international conference on educational data mining*.

Rad, N. M. and Furlanello, C. (2016). Applying deep learning to stereotypical motor movement detection in autism spectrum disorders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 1235–1242. IEEE.

Rajagopalan, S., Dhall, A., and Goecke, R. (2013). Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 755–761.

Rajagopalan, S. S., Morency, L.-P., Baltrusaitis, T., and Goecke, R. (2016). Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, pages 338–353. Springer.

Rajagopalan, S. S., Murthy, O. R., Goecke, R., and Rozga, A. (2015). Play with me—Measuring a child's engagement in a social interaction. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE.

Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., and Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3053.

Raptis, M. and Sigal, L. (2013). Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657.

Rasalingam, R.-R., Muniandy, B., and Rass, R. (2014). Exploring the application of Augmented Reality technology in early childhood classroom in Malaysia. *Journal of Research & Method in Education (IOSR-JRME)*, 4(5):33–40.

Rasipuram, S. and Jayagopi, D. B. (2019). A comprehensive evaluation of audio-visual behavior in various modes of interviews in the wild. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 94–100. ACM.

Reddy, A. M. and Raj, B. (2007). Soft mask methods for single-channel speaker separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1766–1776.

Regina Molini-Avejonas, D., Rondon-Melo, S., de La Higuera Amato, C. A., and Samelli, A. G. (2015). A systematic review of the use of telehealth in speech, language and hearing sciences. *Journal of Telemedicine and Telecare*, 21(7):367–376.

Rehg, J. M., Rozga, A., Abowd, G. D., and Goodwin, M. S. (2014). Behavioral imaging and autism. *IEEE Pervasive Computing*, 13(2):84–87.

Reyes, A. K., Caicedo, J. C., and Camargo, J. E. (2015). Fine-tuning Deep Convolutional Networks for Plant Recognition. *CLEF (Working Notes)*, 1391.

Robinson, S. E. (2011). Teaching paraprofessionals of students with autism to implement pivotal response treatment in inclusive school settings using a brief video feedback training package. *Focus on Autism and Other Developmental Disabilities*, 26(2):105–118.

Robles-Bykbaev, V. E., López-Nores, M., Pazos-Arias, J. J., and Arévalo-Lucero, D. (2015). SPELTA: An expert system to generate therapy plans for speech and language disorders. *Expert Systems with Applications*, 42(21):7641–7651.

Rüping, S. (2004). A simple method for estimating conditional probabilities for svms. Technical report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in ....

Rudovic, O., Lee, J., Dai, M., Schuller, B., and Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760.

Ryoo, M. S. and Aggarwal, J. (2010). UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). In *IEEE International Conference on Pattern Recognition Workshops*, volume 2, page 4.

Sadjadi, S. O. and Hansen, J. H. (2013). Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing Letters*, 20(3):197–200.

Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4580–4584. IEEE.

Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.

Salehi, H. and Parsa, V. (2016). Nonintrusive speech quality estimation based on Perceptual Linear Prediction. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 305–311. IEEE.

Sarikaya, R., Hinton, G. E., and Deoras, A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):778–784.

Sarokoff, R. A. and Sturmey, P. (2004). The effects of behavioral skills training on staff implementation of discrete-trial teaching. *Journal of Applied Behavior Analysis*, 37(4):535–538.

Sarokoff, R. A. and Sturmey, P. (2008). The effects of instructions, rehearsal, modeling, and feedback on acquisition and generalization of staff use of discrete trial teaching and student correct responses. *Research in Autism spectrum disorders*, 2(1):125–136.

Scahill, L., Riddle, M. A., McSwiggin-Hardin, M., Ort, S. I., King, R. A., Goodman, W. K., Cicchetti, D., and Leckman, J. F. (1997). Children's Yale-Brown obsessive compulsive scale: reliability and validity. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(6):844–852.

Schimmel, S. M., Atlas, L. E., and Nie, K. (2007). Feasibility of single channel speaker separation based on modulation frequency analysis. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–605. IEEE.

Schindler, K. and Van Gool, L. J. (2008). Action snippets: How many frames does human action recognition require? In *CVPR*, volume 1, pages 3–2.

Schreibman, L., Dawson, G., Stahmer, A. C., Landa, R., Rogers, S. J., McGee, G. G., Kasari, C., Ingersoll, B., Kaiser, A. P., Bruinsma, Y., and others (2015). Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 45(8):2411–2428.

Schreibman, L., Kaneko, W. M., and Koegel, R. L. (1991). Positive affect of parents of autistic children: A comparison across two teaching techniques. *Behavior Therapy*, 22(4):479–490.

Schuller, B. W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A. C., Zhang, Y., Coutinho, E., and Evanini, K. (2016). The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Interspeech*, pages 2001–2005.

Sener, F. and Ikizler-Cinbis, N. (2015). Two-person interaction recognition via spatial multiple instance embedding. *Journal of Visual Communication and Image Representation*, 32:63–73.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). NTU RGB+ D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019.

Sheena, C. V. and Narayanan, N. (2015). Key-frame extraction by analysis of histograms of video frames using statistical methods. *Procedia Computer Science*, 70:36–40.

Sherer, M. R. and Schreibman, L. (2005). Individual behavioral profiles and predictors of treatment effectiveness for children with autism. *Journal of consulting and clinical psychology*, 73(3):525.

Shete, D., Patil, S., and Patil, S. (2014). Zero crossing rate and Energy of the Speech Signal of Devanagari Script. *IOSR-JVSP*, 4(1):1–5.

Shin, J. W., Chang, J.-H., and Kim, N. S. (2010). Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530.

Shin, M., Bryant, D. P., Bryant, B. R., McKenna, J. W., Hou, F., and Ok, M. W. (2016). Virtual Manipulatives Tools for Teaching Mathematics to Students With Learning Disabilities. *Intervention in School and Clinic*, page 1053451216644830.

Shivakumar, P. G., Li, H., Knight, K., and Georgiou, P. (2018). Learning from Past Mistakes: Improving Automatic Speech Recognition Output via Noisy-Clean Phrase Context Modeling. *arXiv preprint arXiv:1802.02607*.

Shivakumar, S. S., Loeb, H., Bogen, D. K., Shofer, F., Bryant, P., Prosser, L., and Johnson, M. J. (2017). Stereo 3d tracking of infants in natural play conditions. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pages 841–846. IEEE.

Shrawankar, U. and Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.

Signh, N. (2014). *The Effects of Parent Training in Pivotal Response Treatment (PRT) and Continued Support through Telemedicine on Gains in Communication in Children with Autism Spectrum Disorder*. Degree of Doctor of Medicine, University of Arizona.

Skokut, M., Robinson, S., Openden, D., and Jimerson, S. R. (2008). Promoting the social and cognitive competence of children with autism: Interventions at school. *The California School Psychologist*, 13(1):93–108.

Smith, D., Sneddon, A., Ward, L., Duenser, A., Freyne, J., Silvera-Tawil, D., and Morgan, A. (2017). Improving child speech disorder assessment by incorporating out-of-domain adult speech. *Proc. Interspeech 2017*, pages 2690–2694.

Smith, I. M., Flanagan, H. E., Garon, N., and Bryson, S. E. (2015). Effectiveness of community-based early intervention based on pivotal response treatment. *Journal of Autism and Developmental Disorders*, 45(6):1858–1872.

Smith, T. (2001). Discrete trial training in the treatment of autism. *Focus on autism and other developmental disabilities*, 16(2):86–92.

Smith, T. (2010). Early and intensive behavioral intervention in autism.

Solutions, B. I. (2018a). Behavior Connect – Behavior ImagingRetrieved from: https://behaviorimaging.com/products/behavior-connect/.

Solutions, B. I. (2018b). Behavior Imaging – Health & Education Assessment Technology, Retrieved from: https://behaviorimaging.com/.

Sparrow, S. S., Cicchetti, D. V., and Balla, D. A. (1989). The vineland adaptive behavior scales. *Major psychological assessment instruments*, 2:199–231.

Specia, L., Saunders, C., Turchi, M., Wang, Z., and Shawe-Taylor, J. (2009). Improving the confidence of machine translation quality estimates. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 136–143.

Sprafkin, J., Volpe, R. J., Gadow, K. D., Nolan, E. E., and Kelly, K. (2002). A DSM-IV–referenced screening instrument for preschool children: The Early Childhood Inventory-4. *Journal of the American Academy of Child & Adolescent Psychiatry*, 41(5):604–612.

Stahmer, A. C. (1995). Teaching symbolic play skills to children with autism using pivotal response training. *Journal of autism and developmental disorders*, 25(2):123–141.

Stahmer, A. C., Schreibman, L., and Powell, N. P. (2006). Social validation of symbolic play training for children with autism. *Journal of Early and Intensive Behavior Intervention*, 3(2):196.

Steiner, A. M., Gengoux, G. W., Klin, A., and Chawarska, K. (2013). Pivotal response treatment for infants at-risk for autism spectrum disorders: A pilot study. *Journal of autism and developmental disorders*, 43(1):91–102.

Sturmey, P. and Fitzer, A. (2007). *Autism spectrum disorders: Applied behavior analysis, evidence, and practice*. Pro-ed.

Subramanya, A., Srinivas, S., and Babu, R. V. (2017). Confidence estimation in deep neural networks via density modelling. *arXiv preprint arXiv:1707.07013*.

Suhrheinrich, J. and Chan, J. (2017). Exploring the Effect of Immediate Video Feedback on Coaching. *Journal of Special Education Technology*, 32(1):47–53.

Suhrheinrich, J., Reed, S., Schreibman, L., and Bolduc, C. (2011). *Classroom pivotal response teaching for children with autism*. Guilford Press.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. (2014). Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., and Griss, M. (2010). Activity-aware mental stress detection using physiological sensors. In *International conference on Mobile computing, applications, and services*, pages 282–301. Springer.

Sun, M. and Savarese, S. (2011). Articulated part-based model for joint object detection and pose estimation. In *2011 International Conference on Computer Vision*, pages 723–730. IEEE.

Suomalainen, M. and Kyrki, V. (2017). A geometric approach for learning compliant motions from demonstration. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 783–790. IEEE.

Symon, J. B. (2005). Expanding interventions for children with autism: Parents as trainers. *Journal of Positive Behavior Interventions*, 7(3):159–173.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tadayon, R., Amresh, A., McDaniel, T., and Panchanathan, S. (2018). Real-time stealth intervention for motor learning using player flow-state. In *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–8. IEEE.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Tamura, Y., Akashi, T., Yano, S., and Osumi, H. (2016). Human Visual Attention Model Based on Analysis of Magic for Smooth Human–Robot Interaction. *International Journal of Social Robotics*, 8(5):685–694.

Tamura, Y., Yano, S., and Osumi, H. (2014). Modeling of human attention based on analysis of magic. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 302–303. ACM.

Tanenbaum, A. S. and Van Steen, M. (2007). *Distributed systems: principles and paradigms*. Prentice-Hall.

Thakre, K., Rajurkar, A., and Manthalkar, R. (2016). Video partitioning and secured keyframe extraction of MPEG video. *Procedia Computer Science*, 78:790–798.

Thanda, A. and Venkatesan, S. M. (2017). Multi-task learning of deep neural networks for audio visual automatic speech recognition. *arXiv preprint arXiv:1701.02477*.

Thorp, D. M., Stahmer, A. C., and Schreibman, L. (1995). Effects of sociodramatic play training on children with autism. *Journal of autism and developmental disorders*, 25(3):265–282.

Tian, Y., Zitnick, C. L., and Narasimhan, S. G. (2012). Exploring the spatial hierarchy of mixture models for human pose estimation. In *European Conference on Computer Vision*, pages 256–269. Springer.

Titze, I. R. and Martin, D. W. (1998). *Principles of voice production*. ASA.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660.

Tripathi, S., Lipton, Z. C., Belongie, S., and Nguyen, T. (2016). Context matters: Refining object detection in video with recurrent neural networks. *arXiv preprint arXiv:1607.04648*.

Tsatsoulis, P. D., Kordas, P., Marshall, M., Forsyth, D., and Rozga, A. (2016). The Static Multimodal Dyadic Behavior Dataset for Engagement Prediction. In *Computer Vision–ECCV 2016 Workshops*, pages 386–399. Springer.

Tuske, Z., Golik, P., Schlüter, R., and Ney, H. (2014). Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Fifteenth annual conference of the international speech communication association*.

Urbieta, M., Torres, N., Rivero, J. M., Rossi, G., and Dominguez-Mayo, F. J. (2018). Improving Mockup-Based Requirement Specification with End-User Annotations. In *International Conference on Agile Software Development*, pages 19–34. Springer, Cham.

Van Gemeren, C., Poppe, R., and Veltkamp, R. C. (2016). Spatio-Temporal Detection of Fine-Grained Dyadic Human Interactions. In *International Workshop on Human Behavior Understanding*, pages 116–133. Springer.

Venkateswara, H., McDaniel, T., Tadayon, R., and Panchanathan, S. (2018). Person-Centered Technologies for Individuals with Disabilities: Empowerment Through Assistive and Rehabilitative Solutions. *Technology & Innovation*, 20(1-2):117–132.

Ventola, P., Friedman, H. E., Anderson, L. C., Wolf, J. M., Oosting, D., Foss-Feig, J., McDonald, N., Volkmar, F., and Pelphrey, K. A. (2014). Improvements in social and adaptive functioning following short-duration PRT program: a clinical replication. *Journal of autism and developmental disorders*, 44(11):2862–2870.

Verschuur, R., Huskens, B., and Didden, R. (2019). Effectiveness of Parent Education in Pivotal Response Treatment on Pivotal and Collateral Responses. *Journal of autism and developmental disorders*, pages 1–17.

Virgir05 (2015). Pivotal Response Treatment PRT example.

Vismara, L. A. and Lyons, G. L. (2007). Using perseverative interests to elicit joint attention behaviors in young children with autism: Theoretical and clinical implications for understanding motivation. *Journal of Positive Behavior Interventions*, 9(4):214–228.

Vismara, L. A., McCormick, C., Young, G. S., Nadhan, A., and Monlux, K. (2013). Preliminary findings of a telehealth approach to parent training in autism. *Journal of Autism and Developmental Disorders*, 43(12):2953–2969.

Vismara, L. A., Young, G. S., and Rogers, S. J. (2012). Telehealth for expanding the reach of early autism training to parents. *Autism research and treatment*, 2012.

Vismara, L. A., Young, G. S., Stahmer, A. C., Griffith, E. M., and Rogers, S. J. (2009). Dissemination of evidence-based practice: Can we train therapists from a distance? *Journal of autism and developmental disorders*, 39(12):1636.

Voulodimos, A., Doulamis, N., Doulamis, A., Lalos, C., and Stentoumis, C. (2016). Human tracking driven activity recognition in video streams. In *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*, pages 554–559. IEEE.

Vázquez-Martín, R. and Bandera, A. (2013). Spatio-temporal feature-based keyframe detection from video shots using spectral clustering. *Pattern Recognition Letters*, 34(7):770–779.

Wang, Q.-F., Yin, F., and Liu, C.-L. (2011). Improving handwritten Chinese text recognition by confidence transformation. In *2011 International Conference on Document Analysis and Recognition*, pages 518–522. IEEE.

Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480.

Ward, L., Stefani, A., Smith, D., Duenser, A., Freyne, J., Dodd, B., and Morgan, A. (2016). Automated screening of speech development issues in children by identifying phonological error patterns. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, pages 2661–2665.

WaveSurfer (2018). WaveSurfer User Manual, Retrieved from: https://www.speech.kth.se/wavesurfer/man.html.

Wei, P., Xie, D., Zheng, N., and Zhu, S.-C. (2017). Inferring Human Attention by Learning Latent Intentions. *Procedings of the Twenth-Sixth International Joint Conference of Artificial Intellegence*.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Wilpon, J. G., Rabiner, L. R., Lee, C.-H., and Goldman, E. (1990). Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878.

Winne, P. H. (2011). A Cognitive and Metacognitive Analysis of Self-Regulated Learning: Faculty of Education, Simon Fraser University, Burnaby, Canada. In *Handbook of self-regulation of learning and performance.*, pages 29–46. Routledge.

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., Brock, M. E., Plavnick, J. B., Fleury, V. P., and Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders*, 45(7):1951–1966.

Xia, R. and Liu, Y. (2017). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1):3–14.

Xu, D., Gilkerson, J., Richards, J., Yapanel, U., and Gray, S. (2009). 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*, pages 2518–2522. IEEE.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., and Hansen, J. (2008). Signal processing for young child speech language development. In *First Workshop on Child, Computer and Interaction*.

Xu, J., Song, L., Xu, J. Y., Pottie, G. J., and Van Der Schaar, M. (2016). Personalized active learning for activity classification using wireless wearable sensors. *IEEE journal of selected topics in signal processing*, 10(5):865–876.

Yilmaz, R. M. (2016). Educational magic toys developed with augmented reality technology for early childhood education. *Computers in Human Behavior*, 54:240–248.

Young, J. M., Krantz, P. J., McClannahan, L. E., and Poulson, C. L. (1994). Generalized imitation and response-class formation in children with autism. *Journal of Applied Behavior Analysis*, 27(4):685–697.

Yu, D., Kolba ek, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.

Zeng, M., Gao, H., Yu, T., Mengshoel, O. J., Langseth, H., Lane, I., and Liu, X. (2018). Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 56–63. ACM.

Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. (2016). Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer.

Zhang, X.-L. and Wang, D. (2016). Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(2):252–264.

Zhang, Y., Liu, X., Chang, M.-C., Ge, W., and Chen, T. (2012). Spatio-temporal phrases for activity recognition. In *European Conference on Computer Vision*, pages 707–721. Springer.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., and Courville, A. (2017a). Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.

Zhang, Y., Yan, D., and Yuan, Y. (2017b). An object tracking algorithm with embedded gyro information. In *Seventh International Conference on Electronics and Information Engineering*, volume 10322, page 103220U. International Society for Optics and Photonics.

Zhao, L., Sukthankar, G., and Sukthankar, R. (2011). Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 728–733. IEEE.

Zhao, R., Ali, H., and van der Smagt, P. (2017). Two-Stream RNN/CNN for Action Recognition in 3d Videos. *arXiv preprint arXiv:1703.09783*.

Zimmerman, I., Steiner, V., and Pond, R. (2002). *Preschool Language Scale, Fourth Edition*. The Psychological Corporation, San Antonio, TX.