Using Multimodal Data for Automated Fidelity Evaluation in Pivotal Response Treatment Videos

Corey DC Heath, Hemanth Venkateswara, Troy McDaniel, Sethuraman Panchanathan
Center of Cognitive Ubiquitous Computing
Arizona State University
Tempe, AZ, USA
{corey.heath,hemanthv,troy.mcdaniel,panch}@asu.edu

***Abstract*—Research has shown that caregivers implementing pivotal response treatment (PRT) with their child with autism spectrum disorder (ASD) helps the child develop social and communication skills. Evaluation of caregiver fidelity to PRT in training programs and research studies relies on the evaluation of video probes depicting the caregiver interacting with his or her child. These video probes are reviewed by behavior analysts and are dependent on manual processing to extract data metrics. Using multimodal data processing techniques and machine learning could alleviate the human cost of evaluating the video probes by automating data analysis tasks.**

Creating an 'Opportunity to Respond' is one of the categories used to evaluate caregiver fidelity to PRT implementation. A caregiver is determined to have successfully demonstrated creating an opportunity to respond when they have delivered an appropriate instruction while she or he has the child's attention. Automatically determining when the caregiver has correctly provided an opportunity to respond requires classifying the audio and video data from the probes. Combining the modalities into a single classification task can be undertaken using feature fusion or decision fusion methods. Two decision fusion configurations, and a feature fusion model were evaluated. The decision fusion models achieved higher accuracy, however the feature fusion model had a higher average F1 score, indicating more reliable prediction capability.

***Index Terms*—Attention Detection, Pivotal Response Treatment, Multimodal Data, Autism Spectrum Disorder, Machine Learning**

I. Introduction

Pivotal response treatment (PRT) is an applied behavior analysis (ABA) technique that focuses on presenting learning objectives in a natural context [1]. Research involving PRT has primarily explored using the technique to aid children with autism spectrum disorder (ASD) in developing social and communication skills [2]–[4]. For these sessions, the interventionist observes the child to determine an activity the child is interested in engaging in. The interventionist can capitalize on the child's natural motivation to continue the desired activity to ensure compliance with learning objectives. To do this, the interventionist needs to be able to gain control of the activity to be able to integrate learning objects into it, or stop the activity to elicit the child's attention. Once the interventionist has the child's attention, he or she should deliver an instruction that is at an appropriate language level for the child. After the child response, the intervention will continue the motivating activity if the child has demonstrated a legitimate attempt at the prompted skill. If the child did not demonstrate an appropriate attempt, the interventionist should hold the activity contingent on adequate compliance. In PRT, the process of gaining the child's attention and delivering an instruction is often referred to as creating an 'Opportunity to Respond'.

Evaluation of the caregiver's performance as the interventionist in PRT is often assessed using video probes of the caregivers interacting with their child [5]–[7]. These videos are scored based on four meta-categories including identifying the natural reinforcer for the child, creating an opportunity to respond, varying instructions and targeted skills, and being contingent on an appropriate response from the child. The opportunity to respond category is composed of gaining the child's attention, and issuing a clear instruction at the child's language level. To assess these categories, the clinician divides the video into one or two minute increments and provides a binary score for each subsection. The caregiver is expected to provide at least two opportunities to respond per minute to gain a positive score for that interval.

Automatically assessing if the caregiver has provided an opportunity to respond depends on detecting if the child is attentive to the caregiver, and if the caregiver vocalizes during this period of attention. Determining the attention state of the child utilizes computer vision techniques to classify the visual cues of attention. Likewise, evaluating the caregiver's instruction involves identifying the adult speaker in the audio.

The research presented below will focus on detecting segments of PRT video probes that could contain opportunities to respond. This is defined by the child being attentive or in a joint attentive state with the caregiver while the caregiver is speaking. For this work, evaluating the language usage in the adult instruction will not be included.

There are two basic methodologies for detecting candidate segments for opportunities to respond that will be employed. First, using decision fusion, the audio and visual modalities will be classified separately, and the predictions will be combined to infer if an opportunity to respond has occurred. Secondly, using feature fusion, the features from the modalities will be combined and used to train a model directly on a binary classification task based on whether or not a sample is a candidate for an opportunity to respond.

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

II. Related Work

The problem addressed in this work required examining human behavior in video image data and detecting adult vocalizations from an audio track. The approach undertaken for video processing involved extracting spatio-temporal (ST) graphs of the individuals in the video frames [8]. Audio processing aims at performing voice activity detection (VAD) and speaker separation. These separate modalities are fused to classify intervention intervals from the dataset.

ST graphs are a common method of feature extraction for machine learning models for detecting human activity from video data. The graph represents a skeletal representation of the figure through the frames, with the graph nodes representing major body landmark or articulation points. This has been used for activity detection and motion casting using machine and deep learning methodologies, including support vector machines (SVM) [9] and recurrent neural networks (RNN) [10], [11]. Extending ST graphs for multiple person interactions has been explored in [12]–[14].

Detecting joint attention and social engagement in videos has focused on gaze estimation and proximity [15], [16]. Engagement between a child and clinician were examined by [17], [18]. These sessions focused on pre-specified activities, and utilized multiple stationary camera perspectives.

VAD consists of techniques for discerning human vocalization from other sounds in an audio track. Commonly employed techniques focus on discriminatory features, such as frequency [19], [20], and statistical approaches focusing on modeling noise signals [21]. SVM classification models have been used for binary VAD [22]–[24]. Dyadic speech classification for children and adults has focused on domain adaptation [25].

Multimodal classification involves incorporating data from different sources into a single model. Incorporating the different modalities can occur at different places in the classification process. Feature fusion, or early feature fusion, refers to instances where the data from the modalities is merged prior to training the classifier. Decision fusion, or late feature fusion, involves training classification models on each modality independently before merging the classifier output to infer a final prediction. The classifier output could be a transformed representation of the features, the soft-max probabilities for different classes, or a predicted label. Merging the output could utilize various methods, including additional classification models [26]. Research involving multimodal models using audio and visual data have focused on speech recognition [27], speech prediction [28], and video descriptions [29].

The application of audio and visual machine learning for children with autism has largely focused on detection, diagnosis, and emotion recognition. Regarding activity detection in video data, identifying self-stimulating behaviors has been examined [30]. Exploring speech-based emotion recognition, Marchi et al. [31] trained a classification model to detect a child's affective state based on a story prompt. Their study focused on a child diagnosed with ASD, and a child without. The LENA recording system was developed to analyze vocal data for children with autism. The system focuses on evaluating language skills and social interactions [32]–[34]. A multimodal audio-visual approach to affect and engagement classification for children with autism was presented in [35]. For this approach, the researchers used feature fusion to train a classification model model using audio, visual, and electrodermal signal data in addition to contextual information.

III. Methodology

The dataset for the project consists of a baseline and a post-treatment video for seven caregiver-child dyads, giving 14 videos total. The videos are approximately 10 minutes long, and depict the caregiver and child engaging in varied activities including playing with toys, watching a movie on a mobile phone, and spinning in a chair. The videos illustrate a challenging, 'in-the-wild' dataset. Each video was filmed with a hand-held device and include periods of instability, and full or partial occlusion of either the caregiver or child. Thirty-frame segments, representing one second of video, were labeled based on three states of attention: attentive, inattentive, and shared attention. The child was considered attentive if they were actively focused on the caregiver. The segment was labeled as inattentive when the child was engaged in a solitary activity, such as playing with a toy or moving about the room. The shared attention state was indicated when the caregiver and the child were engaged in a joint activity, such as playing a game or watching a video. More information regarding video processing on the dataset is presented in [36].

For evaluating attention, features were extracted from the videos using OpenPose [37] as described in [36]. This process involved extracting body articulation points and facial landmarks for the individual identified in each frame of the video. Using the points, along with inferred information, such as gaze estimation, a ST graph of the dyad was constructed for the caregiver-child dyad.

The audio track also posed a challenge for classification tasks. The activities and toys recorded in the video would obscure caregiver and child vocalizations. When not obscured, the caregiver vocalizations could be difficult to detect due to the use of child-directed speech patterns. Additionally, the children depicted in the videos exhibited varying vocal communication abilities. As attempts at words consisting of only phonemes are acceptable in vocalizations in PRT, every vocal sound made by the child should be classified appropriately. The audio data was examined in detail in [38]. The dataset was labeled as adult speech, child vocalization, or noise at 250 ms intervals. Each interval was converted to a 68 element vector including midterm features using pyAudioAnalysis [39]. 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.

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 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. Table I displays the number of opportunity to respond candidate segments identified in each validation set in the dataset.

Three experiments were run to detect opportunity to respond segments - two decision fusion methods and one feature fusion method. Each method used the SVM implementation provided by [40]. The SVM used a C value of 10, a gamma value of 0.01, and a radial basis function kernel. First, the predictions from three class classification models for attention and speaker separation where aggregated to infer when a segment was a candidate for an opportunity to respond. The classification models were based on [36], [38]. Next, two class models were trained on attention and speaker separation separately. For these models, the attentive and shared attention labels were merged to form the true class for the attention classifier. Similarly, the child and noise labels for the audio data were merged to form the negative class to train a model for selecting adult speech samples. The average accuracy for the classification models of the decision fusion methods is displayed in table II. For these two methods, an opportunity to respond label was predicted when the attention state was attentive or shared, and the corresponding audio segment indicated adult speech.The final experiment trained a model directly on samples with an opportunity to respond label. This classifier was trained using a concatenated vector consisting of both the video and audio features as the input. This vector had 133 elements consisting of 68 audio features and 65 video features.

Each of the three approaches were validated using a 'leave-one-dyad-out' method for creating a test set. The data from the base and post-videos for a single caregiver-child dyad were retained to validate the classification models. The remaining 12 videos were used to form a training set. Prior to training the model, the training set was randomized and balanced by randomly undersampling over represented classes.

TABLE I
NUMBER OF OPPORTUNITY TO RESPOND (OTR) 250MS SEGMENTS FOR EACH DYAD IN THE DATASET.

Dyad	***OTR Seg***	***Total Seg***	***OTR Seg Time (sec)***
Dyad1	336	2982	84.00
Dyad2	642	4022	160.50
Dyad3	1043	4141	260.75
Dyad4	691	3445	172.75
Dyad5	524	4716	131.00
Dyad6	452	3968	113.00
Dyad7	834	3846	208.50

IV. RESULTS AND DISCUSSION

The results from detecting opportunity to respond candidate samples from the data set are displayed in table III. The decision fusion accuracy was similar for both the two and three class models, averaging 79% and 80% respectively. 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 below fifty percent for both decision fusion methods. This illustrates the classifier could not adequately distinguish when true samples were present.

TABLE II
AVERAGE VALIDATION ACCURACY OF TWO AND THREE CLASS ATTENTION AND AUDIO-SPEAKER SEPARATION SVM MODELS.

	2-C Attn	***3-C Attn***	***2-C Audio***	***3-C Audio***
Accuracy	0.55	0.43	0.81	0.72

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 over the attention models used in decision fusion methods dominated the aggregated 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 video data used for the attention classification. This indicates that the audio features may be useful in improving the attention classification.

The results for dyad4 were an outlier among the validation sets, having an accuracy score roughly ten 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 in 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 dyad4 validation set at the same performance level as the other sets.

TABLE III
ACCURACY AND F1 SCORES FOR CLASSING OPPORTUNITY TO RESPOND. THE TABLE DISPLAYS RESULTS FOR DECISION FUSION USING TWO AND THREE CLASS MODELS, AND FEATURE FUSION.

	2-Class Dec. Fusion			3-Class Dec. Fusion			Feature Fusion		
Dyad	***Accuracy***	***F1 True***	***F1 False***	***Accuracy***	***F1 True***	***F1 False***	***Accuracy***	***F1 True***	***F1 False***
Dyad1	0.83	0.43	0.90	0.83	0.40	0.90	0.73	0.74	0.74
Dyad2	0.82	0.49	0.89	0.82	0.48	0.89	0.74	0.76	0.79
Dyad3	0.75	0.47	0.84	0.75	0.45	0.84	0.68	0.70	0.71
Dyad4	0.62	0.40	0.73	0.67	0.35	0.78	0.62	0.53	0.59
Dyad5	0.84	0.45	0.90	0.84	0.44	0.91	0.72	0.74	0.75
Dyad6	0.90	0.50	0.94	0.90	0.44	0.94	0.72	0.76	0.83
Dyad7	0.75	0.52	0.83	0.77	0.52	0.85	0.71	0.69	0.68
Average	0.79	0.47	0.86	0.80	0.44	0.87	0.70	0.70	0.73

V. LIMITATIONS AND OPPORTUNITIES FOR FUTURE WORK

This study was a preliminary investigation on how opportunities to respond could be detected in PRT videos and has several limitations. The classifications in the experiments above do not consider the language used by the caregiver in issuing an instruction. This means that although the caregiver spoke during a period of attention the instruction may not be clear or at the child's ability level, and thus a clinician would not consider it a valid opportunity to respond. Additionally, this methodology in this study did not consider cases where a nonverbal instruction was given.

The experiments classified 250 ms samples. This is too small of a sample size to be valuable to clinicians. For the results to be usable in a real world scenario, techniques need to be explored to aggregate the samples to encompass a meaningful amount of time. This would also be necessary for employing natural language processing techniques to evaluate if the caregiver's instructions are clear and concise.

The SVM parameters were standardized for each of the classification tasks. Fine-tuning the parameters for each task may have resulted in better performance.

From a signal processing perspective, the experiments largely removed the temporal relationship between sample sizes. The temporal relationship between events, body poses, and language composition could be utilized to improve the classification techniques.

VI. CONCLUSION

Exploring options for automatically gathering performance metrics and classifying interactions in PRT videos would alleviate the manual cost of training and supporting caregivers learning to practice ABA intervention methods with their child with ASD. The goal of this publication was to examine multimodal methodologies for identifying video segments that could potentially be scored as an acceptable opportunity to respond based on PRT evaluation methodology. This involved inferring the child's attention state based on visual data, along with identifying adult speech from the video's audio track. The data from both modalities was examined using both decision fusion and feature fusion. Decision fusion on two and three class classification models illustrated a high accuracy and low F1 score for the true class, indicating poor performance on the imbalanced data set. Concatenating the feature vectors for both modalities and training a single classifier produced a higher F1 score for the true class, reflecting that the model was able to identify discernible features for each label class. This work was limited by the small size of the dataset. In the future, the PRT video probe dataset needs to be expanded to incorporate more individuals and scenarios to ensure generalizations of approach. Additional datasets regarding dyadic interactions, such as academic instruction, counseling, and interviews should also be explored to determine how well this approach transfers to similar domains.

VII. ACKNOWLEDGEMENT

The authors thank Arizona State University and the National Science Foundation for their funding support. This material is partially based upon work supported by the National Science Foundation under Grant No. 1069125 and 1828010.

REFERENCES

[1] R. L. Koegel, *How To Teach Pivotal Behaviors to Children with Autism: A Training Manual.*, 1988.

[2] L. K. Koegel, R. L. Koegel, I. Green-Hopkins, and C. C. Barnes, "Brief report: Question-asking and collateral language acquisition in children with autism," *Journal of Autism and Developmental Disorders*, vol. 40, no. 4, pp. 509–515, 2010.

[3] R. L. Koegel, J. L. Bradshaw, K. Ashbaugh, and L. K. Koegel, "Improving question-asking initiations in young children with autism using pivotal response treatment," *Journal of autism and developmental disorders*, vol. 44, no. 4, pp. 816–827, 2014.

[4] P. Ventola, H. E. Friedman, L. C. Anderson, J. M. Wolf, D. Oosting, J. Foss-Feig, N. McDonald, F. Volkmar, and K. A. Pelphrey, "Improvements in social and adaptive functioning following short-duration PRT program: a clinical replication," *Journal of autism and developmental disorders*, vol. 44, no. 11, pp. 2862–2870, 2014.

[5] J. Coolican, I. M. Smith, and S. E. Bryson, "Brief parent training in pivotal response treatment for preschoolers with autism," *Journal of Child Psychology and Psychiatry*, vol. 51, no. 12, pp. 1321–1330, 2010.

[6] A. Y. Hardan, G. W. Gengoux, K. L. Berquist, R. A. Libove, C. M. Ardel, J. Phillips, T. W. Frazier, and M. B. Minjarez, "A randomized controlled trial of pivotal response treatment group for parents of children with autism," *Journal of Child Psychology and Psychiatry*, vol. 56, no. 8, pp. 884–892, 2015.

[7] I. M. Smith, H. E. Flanagan, N. Garon, and S. E. Bryson, "Effectiveness of community-based early intervention based on pivotal response treatment," *Journal of Autism and Developmental Disorders*, vol. 45, no. 6, pp. 1858–1872, 2015.

[8] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1282–1289.

[9] C. Y. Chen and K. Grauman, "Efficient activity detection in untrimmed video with max-subgraph search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 908–921, 2017.

[10] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317.

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[12] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 707–721.

[13] F. Sener and N. Ikizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 63–73, 2015.

[14] C. Van Gemeren, R. Poppe, and R. C. Veltkamp, "Spatio-temporal detection of fine-grained dyadic human interactions," in *International Workshop on Human Behavior Understanding*. Springer, 2016, pp. 116–133.

[15] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, vol. 30, no. 2, pp. 115–127, 2013.

[16] S. Duffner and C. Garcia, "Visual focus of attention estimation with unsupervised incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2264–2272, 2016.

[17] S. S. Rajagopalan, O. R. Murthy, R. Goecke, and A. Rozga, "Play with memeasuring a child's engagement in a social interaction," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[18] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 338–353.

[19] I. V. McLoughlin, "The use of low-frequency ultrasound for voice activity detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[20] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 4, pp. 705–717, 2015.

[21] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 252–264, 2016.

[22] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, vol. 2. IEEE, 2002, pp. 1124–1127.

[23] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.

[24] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[25] M. Kumar, D. Bone, K. McWilliams, S. Williams, T. D. Lyon, and S. Narayanan, "Multi-scale context adaptation for improving child automatic speech recognition in child-adult spoken interactions," *Proc. Interspeech 2017*, pp. 2730–2734, 2017.

[26] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 565–572.

[27] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[28] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3051–3055.

[29] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4193–4202.

[30] M. Jazouli, S. Elhoufi, A. Majda, A. Zarghili, and R. Aalouane, "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*, vol. 10, no. 7, pp. 1270–1274, 2016.

[31] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Blte, P. Arora, and R. Hb-Umbach, "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*, 2015, pp. 115–119.

[32] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *First Workshop on Child, Computer and Interaction*, 2008.

[33] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "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*. IEEE, 2009, pp. 2518–2522.

[34] R. Pawar, A. Albin, U. Gupta, H. Rao, C. Carberry, A. Hamo, R. M. Jones, C. Lord, and M. A. Clements, "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*. IEEE, 2017, pp. 245–248.

[35] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[36] C. D. Heath, H. Venkateswara, T. McDaniel, and S. Panchanathan, "Detecting attention in pivotal response treatment video probes," in *International Conference on Smart Multimedia*, 2018.

[37] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[38] C. D. Heath, T. McDaniel, H. Venkateswara, and S. Panchanathan, "Parent and child voice activity detection in pivotal response treatment video probes," *Human Computer Interaction International*, 2019.

[39] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.