INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Computing Multimodal Dyadic Behaviors during Spontaneous Diagnosis Interviews toward Automatic Categorization of Autism Spectrum Disorder

Chin-Po Chen[1], *Xian-Hong Tseng*[1], *Susan Shur-Fen Gau*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan

`cclee@ee.nthu.edu.tw, gaushufe@ntu.edu.tw`

Abstract

Autism spectrum disorder (ASD) is a highly-prevalent neural developmental disorder often characterized by social communicative deficits and restricted repetitive interest. The heterogeneous nature of ASD in its behavior manifestations encompasses broad syndromes such as, Classical Autism (AD), High-functioning Autism (HFA), and Asperger syndrome (AS). In this work, we compute a variety of multimodal behavior features, including body movements, acoustic characteristics, and turn-taking events dynamics, of the participant, the investigator and the interaction between the two directly from audio-video recordings by leveraging the Autism Diagnostic Observational Schedule (ADOS) as a clinically-valid behavior data elicitation technique. Several of these signal-derived behavioral measures show statistically significant differences among the three syndromes. Our analyses indicate that these features may be pointing to the underlying differences in the behavior characterizations of social functioning between AD, AS, and HFA - corroborating some of the previous literature. Further, our signal-derived behavior measures achieve competitive, sometimes exceeding, recognition accuracies in discriminating between the three syndromes of ASD when compared to investigator's clinical-rating on participant's social and communicative behaviors during ADOS.

Index Terms: behavioral signal processing (BSP), autism spectrum disorder, dyadic interaction, multimodal behaviors

1. Introduction

Recent effort in developing computational framework to better understand socio-communicative aspects of human communication has become a crucial component in the emerging frontier of interdisciplinary research, e.g., social signal processing [1] and behavioral signal processing [2]. One of the key applications is in autism spectrum disorder (ASD). ASD is a highly heterogeneous and highly prevalent (1 in 88 children in the USA [3]) neuro-developmental disorder. Recent studies have started to demonstrate the use of low-level behavioral cues in studies of ASD. For example, Marchi et al. demonstrate that the low level acoustic descriptors characterizing emotion expression of ASD children can be used to differentiate between typically- (TD) and ASD children [4]. Another study also shows that a variety of acoustic features derived from spoken sentences can be used to classify other diseases related to ASD (e.g., TD, Pervasive Developmental Disorder (PDD), Pervasive Developmental Disorder Not Otherwise specified (NOS-PDD)) [5]. Further, Liu et al. propose to recognize autism with promising accuracies by modeling facial expression from images [6]; motor abnormalities have also been investigated by Crippa et al. showing that upper body movement can potentially be useful in identifying motor signature of ASD [7].

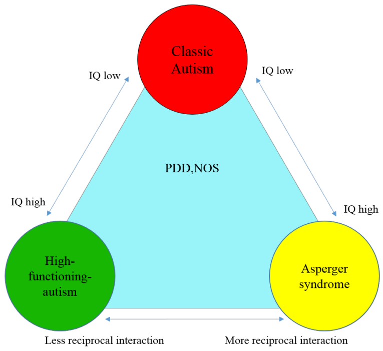

Figure 1: *Schematic of Classical Autism, Asperger syndrome, and High-functioning Autism distribution*

ASD is a combination of disorders that includes previously-defined categories (i.e., AD, HFA, AS) and other mental diseases. HFA was first observed by Kanner and Hans Asperger [8] as a group of individuals standing out to possess above average range on cognitive and linguistic ability. Asperger syndrome was named after Hans Asperger and have been long considered as a similar syndrome to HFA [9, 10]. There has been debates on whether the three syndromes should be viewed as one category of neuro-developmental disorder or be divided into different diseases [10, 11]. Due to synonymousness to autism or otherwise having broad boundaries that may include other mental disorders [12], HFA (autism subjects without cognitive delay [13]) and AS has been merged as ASD in the latest diagnostic tool, Diagnostic and Statistical Manual (DSM-5)[14]. There are, however, still controversies on merging of HFA and AS. Researchers report that AS have similar deficits in social interaction and communication, but also display a rich variety of subtle clinical characteristics that distinguish them from HFA [12, 15]. For example, it is reported that impoverished intonation is a characteristics related more to HFA [16]. In terms of social engagements, AS children display more attention and reciprocal social interactions [17].

In this work, we present a novel study by directly computing behaviors from spontaneous dialogs using audio-video data to quantify the characteristics of deficit among ASD sub-groups in a more granular and quantitative manner, which is a similar concept stated in [18]. Bone et al. have also analyzed the subtle prosodic variations of interlocutors during the Autism Diagnostic Observation Schedule to identify relevant acoustical patterns related to the severity of ASD [19]. ADOS is a standardized assessment for assessing ASD through semi-structured

Thanks to Ministry of Science and Technology (104-2420-H-400-001-MY2) and National Health Research Institute (NHRI-EX104-10404PI, NHRI-EX105-10404PI) for funding

http://dx.doi.org/10.21437/Interspeech.2017-563

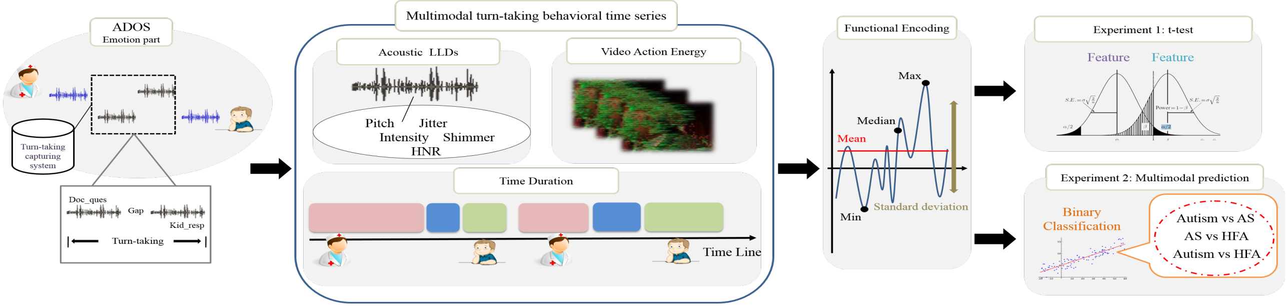

Figure 2: *The architecture of multimodal autism clinical result classification and behavior quantification from audio and video data*

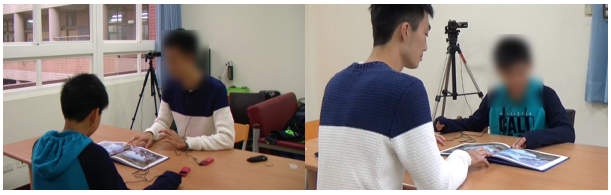

Figure 3: *A demonstration of our experiment setting (a mock up scene resembling the real ADOS collection).*

dyadic interactions in clinical practice [20]. We can conceptualize this procedure of ADOS as composed of two components: 1) *interaction setting*: eliciting natural and targeted behaviors for ASD participants through clinically-valid design of dyadic interactions, 2) *behavior rating*: assessing clinically-relevant behaviors of ASD participants by certified investigator as he/she engages in such an interaction session. By leveraging ADOS interviews as providing clinically-rigorous *interaction settings* for computational behavior analyses, we can measure behaviors, i.e., body movements, vocal characteristics, and turn-taking events dynamics, of the participant, the investigator, and the interaction between the two as alternative signal-derived *behavior ratings* in tasks of analyzing differences among the three different categorizations.

Our analyses reveal that several behavior cues, such as the ratio between normalized body movement energy across a turn-taking event, harnomic-to-noise measure of autistic patient's vocal characteristics, and the response latency during a turn-taking events, show statistically-significance differences between AS and HFA groups. Furthermore, these signal-derived behaviors achieve competitive, sometimes exceeding, recognition accuracies in discriminating between the three categorizations of ASD when compared to investigator's rating on social and communicative ratings (summative and holistic behavior features of the participant during ADOS interviews).

2. Research Methodology

2.1. Database Description

In this work, we analyze the spontaneous interaction data of ADOS diagnosis interviews [1]. ADOS is a clinically-valid instrument in eliciting natural and targeted social communicative behaviors of the participant through semi-structured dyadic interactions. ADOS manual is one of the gold standards in diagnosing and assessing the severity of ASD in clinical practices. The structure of ADOS consists of a series of activities in order for the investigator to elicit and rate the behaviors of the participant related to communication, social interaction, play, imaginative use of materials, etc. Psychiatrists need to go through intensive and rigorous training to become certified in order to be eligible in carrying out an ADOS diagnosis interview.

Table 1: *Demographics of ASD participants in our dataset*

	ASD Clinical Diagnosis		
	Autism (13)	AS (11)	HFA (10)
Age (Avg/Std)	14.2/3.08	15.1/2.80	17.7/4.29
Module (M3/M4)	11/2	7/4	3/7

Our recordings of ADOS sessions, lasts about forty minutes to an hour, are administered at the Department of Psychiatry, National Taiwan University Hospital (NTUH). We set up three high-definition cameras (two of them facing the participant with one facing the investigator), and two lapel microphones (each is clipped on an individual speaker's collar). Figure 3 shows a snapshot of our data collection process from two of the three camcorders. To date, we have collected 34 ASD participants' data during ADOS interviews. The investigator also provides the final clinical diagnoses of these 34 participants based on a combination of ADOS, Autism Diagnostic Interview-Revised (ADIR) [21], and other relevant clinical instrument; 13 of them are assigned as AD, 11 of them are AS, and 10 of them are HFA. Table 1 shows the demographics of our subjects.

2.2. Multimodal Behavior Measurements

Our feature computation involves several major components: segments in turn-taking events, low-level descriptors(LLDs), types of perspectives, segment-level functions, session-level statistics (a schematic flow is shown in Figure 2).

2.2.1. Turn-taking Event and Segmentation

The ability to maintain a smooth conversation require appropriate signaling and reacting during floor exchange, in fact, turn-taking deficit has been demonstrated in the ASD population [22]. Since the ADOS interview involves back-and-forth spoken interactions, we first identify a *turn-taking event* within the conversation as a complete floor exchange between the investigator and the participant. Within each turn-taking event, we split this region into three segments: Invest$_{quest}$, Gap, Part$_{resp}$. Invest$_{quest}$ (the term participant and investigator is abbreviated as "part" and "invest" respectively) defines the region starting from the beginning of investigator's question to the end before giving the turn to the participant. Gap defines the response latency between the end of a question and the beginning of the participant's answer. Finally Part$_{resp}$ defines the entire region where participant answers the investigator's question.

The identification of a *turn-taking event* and subsequently three different *segments* are done both manually and automatically. We leverage our recording setup where we have two channels of audio, each from a speaker in the dialog. We perform speaker identification and segmentation by combining a GMM-based voice activity detector along with algorithm based on energy differential between the two channels. A smoothing algorithm is further applied to filter out "non turn-taking"

[1] Approved by IRB: REC-10501HE002 and RINC-20140319

Table 2: *A list of behavior features and the components required to compute the session-level features. (for convenience, the term participant is abbreviated as "part" and investigator is abbreviated as "invest")*

Component	Items
Segment regions:	Invest$_{quest}$, Gap, Part$_{resp}$
Multimodal LLDs:	*Video*: NBAE, *Audio*: Intensity, Pitch, HNR, Jitter, Shimmer, *Time*: Duration
Perspectives types:	*Inter*-segment (Invest/Part), *Intra*-segment of Invest and Part
Segment-level functions:	*Standard deviation*: Pitch, HNR, *Mean*: Intensity, NBAE
Session-level statistics:	min, max, mean, median, std

events. The manually-defined segments are used in the analysis experiment (section 3.2), and the recognition results for both segmentation methods are provided in section 3.3.

2.2.2. Segment-level Multimodal Behavior Features

We calculate features of three major modalities at *segment*-level for each turn-taking event: video, audio, and time. In the video modality, we compute normalized body action energy features (NBAE) describing relative amount of movement for a person. NBAE is build upon dense trajectory feature extraction [23]. Dense trajectory methods are used to identify the moving objects within a video sequence by tracking the movements of feature points. First, dense feature points (x_t, y_t) are first sampled on a grid space using $W = 5$ pixels and tracked in eight different spatial scales per frame t. Each point $P_t = (x_t, y_t)$ is tracked to the next frame $t + 1$ by median filtering in a dense optical flow field $\omega = (u_t, v_t)$.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(x_{\bar{t}}, y_{\bar{t}})} \quad (1)$$

These tracked points can be thought as an proxy to the amount of movement(equals to the number of moving trajectories of a frame in a video sequence over an interval). We then compute the average number of points being tracked every 15 frames ($\approx$ 0.5 seconds) to be the amount of movement, termed *action energy* (AE). Each AE is extracted at 0.5 second framerate. Deficits of prosody have been described as an integral part of disorder in ASD [24]. Hence, we further compute various low-level prosodic descriptors (LLDs), including pitch, intensity, harmonic-to-noise ratio (HNR), jitter, and shimmer using the Praat toolkit [25]. Pitch, intensity, and HNR are extracted at 10ms per second; these LLDs are further z-normalized with respect to an individual speaker.

Since acoustic LLDs and AEs are at frame level, to derive *segment*-level features, we further compute standard deviation of pitch and HNR and average of intensity and AE over the segments within a turn-taking event (section 2.2.1). Segment-level AEs are z-normalized with respect to individual speaker to derive the NBAE. We additionally include the time aspect, i.e., durational features, as another set of segment-level features. Lastly, aside from computing features within each segment (i.e., termed as *intra*-segment) separately, in order to capture the dynamics between the interlocutors, we further calculate features of *inter*-segment by taking the ratio between the features derived from the investigator's intra-segment and the participant's intra-segment. For example, inter-segment features of Std$(HNR)_{\text{inter}}$ is equal to Std$(HNR)_{\text{Invest-quest}}$/Std$(HNR)_{\text{Part-resp}}$.

2.2.3. Session-level Multimodal Behavior Features

Segment-level multimodal behavior features are computed for every turn-taking event. We further calculate five additional functionals (min, max, mean, median, std) to represent the behaviors at the entire ADOS *session*-level, where the label of each participant is given. For example, *median*-Std$(HNR_{\text{intra-part}})$ is a feature computed by first taking the standard deviation of frame-level HNR values within the participant responding segment of each turn-taking region, then subsequently taking the median over the multiple turn-takings events over a session. Table 2 summarizes the various components and items that are used to to generate all of our *segment*-level features and subsequently *session*-level behavior representations.

3. Experimental Setup and Result

3.1. Experimental Setup

In this work, we conduct two different experiments. The first is to analyze the statistical differences between "AD vs. AS", "AS vs. HFA", and "AD vs. HFA" using the computed session-level features described in section 2.2. The features used in this experiment are derived from manual segmentations to provide a clean analysis. Statistical testing using Student's *t*-tests are carried out with significance level set at $\alpha = 0.05$.

We conduct the second experiment as an automatic recognition task using logistic regression trained on the selected session-level features to discriminate between the three clinical diagnoses of the ASD. In this experiments, we present our recognition accuracies on features derived from both manual and automatic segments identification to resemble the real-life scenarios. Results of our behavior measure are further compared to the psychological ratings done by the investigators during the administration of ADOS. This experiment is carried out in a leave-one-participant-out cross-validation, and we use unweighted average recall to measure our performance.

3.2. Statistical Analysis

Table 3 provides a list of session-level behavior features that show statistically-significant difference at the level of $\alpha = 0.05$ between any two diagnoses. The directions (greater or smaller than) of the differences between groups are also indicated.

We identify several qualitative observations with this statistical analysis result. In "AS vs. HFA", we show that max-$(NBAE_{\text{intra-part}})$ is significantly higher in HFA participant than in AS participants. A closer investigation suggest that: AS participants have lower values in NBAE features because their interactions with the investigator tend to be smoother with the amount of body movements remain more consistent (normal) throughout the ADOS interaction; unlike HFA-

Table 3: *A list of features are significant at* $\alpha = 0.05$*, and the direction of the difference is also indicated.*

Behavior Features	Significance	Direction
$median$-$(NBAE_{\text{intra-Invest}})$	✓	AS>AD
min-$(Duration_{\text{intra-Invest}})$	✓	AS>AD
max-$(Jitter_{\text{intra-part}})$	✓	AS>AD
$median$-Std$(HNR_{\text{intra-part}})$	✓	AS>AD
max-$(Duration_{\text{inter}})$	✓	AS>HFA
$mean$-$(Duration_{\text{inter}})$	✓	AS>HFA
std-$(Duration_{\text{inter}})$	✓	AS>HFA
max-$(NBAE_{\text{intra-part}})$	✓	AS<HFA
max-$(NBAE_{\text{inter}})$	✓	AD<HFA
max-$(NBAE_{\text{gap}})$	✓	AD>HFA
$median$-$(Duration_{\text{gap}})$	✓	AD<HFA

Table 4: *A summary of multimodal feature recognition result. We use three feature modalities: action energy feature (NBAE$_{Partresp}$), acoustic feature (STD(HNR)$_{Partresp}$), and time duration feature (Duration$_{inter}$).*

Behavior Feature Descriptions	Multimodal Features	Manually-marked / Automatically-defined Segments			
		AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
Action Energy feature:	$NBAE_{\text{intra-part}}$	0.38/0.67	0.71/0.63	0.73/0.69	0.43/0.43
$NBAE_{\text{intra-part}}$ (A)	Std-($HNR_{\text{intra-part}}$)	0.63/0.47	0.37/0.43	0.54/0.61	0.20/0.29
	$Duration_{\text{inter}}$	**0.82**/0.61	0.61/0.57	**0.78**/0.53	**0.61**/0.32
Acoustic feature:	A + H	0.50/0.56	0.66/0.45	0.64/0.73	0.45/0.39
Std-($HNR_{\text{intra-part}}$) (H)	H + D	**0.77**/0.58	0.61/0.57	**0.78**/0.70	0.54/0.43
	A + D	**0.77**/0.62	**0.80**/0.62	**0.78**/**0.79**	**0.61**/0.40
Time duration feature:	A + H + D	0.74/0.58	0.71/0.48	0.69/**0.77**	0.58/0.40
$Duration_{\text{inter}}$ (D)		ADOS Diagnostic Behavior Ratings			
	Communication	0.66	0.71	0.45	0.44
	Social	0.66	0.62	0.62	0.44
	Communication + Social	0.75	0.70	0.54	0.50

participants, where some of them could not carry smooth social interaction and show sudden abrupt or irregular (non-smooth) body movement. Since NBAE is z-scored speaker-dependent amount of movement, it is nature to see some of these *sudden* movements that HFA participants would have resulted in a higher value of NBAE. Furthermore, we also observe that *inter*-segment durational features show an overall higher value in group of AS compared to HFA. Neither the durational feature for the investigator-questioning portion nor the participant-responding portion show significant differences by itself; it is the interaction effect between the two that display a significant difference. This phenomena is likely to be in line with the past literature, which indicates that AS participants are more talkative and show higher motivation in social interaction [26].

In observing the difference in tasks of "AD vs. (HFA or AS)", several of the features appear possibility due to the severe deficit in social communicative ability of AD participants. As an example, the higher language and cognitive ability of HFA (AS) is manifested in the the exchanges between the question raised by the investigator and the answering given by the participant. The interaction often turns out to be more *engaging*. In summary, except for the well-known socio-communicative behavior differences between AD versus HFA/AS groups, we also demonstrate subtle behavior differences exhibited during the ADOS interviews between HFA and AS.

3.3. Multimodal Recognition Experiment

Table 4 summarizes our recognition results calculated using both manually and automatically marked turn-taking segments. We use one type of *segment*-level feature from each modality, i.e., $NBAE_{\text{intra-part}}$, Std-($HNR_{\text{intra-part}}$), and $Duration_{\text{inter}}$, and compute *session*-level functionals as input to our automatic ASD categorization system. The three features are chosen due to the statistical analysis result presented in section 3.2. Since our data is from ADOS interactions, social and communication rating done by the clinicians using the ADOS manual can be viewed as the clinically-valid behavior ratings of the participant. In fact, the final clinical diagnosis often relies on these experts' behavior ratings from the ADOS interview. We can thus compare the automatic recognition results from our multimodal behavior features, which are signal-derived and granular, to the ADOS behavior ratings, which are expert-rated and holistics, in tasks of distinguishing between the three syndromes.

The numbers in bold indicate a higher recognition accuracy achieved compared to using ADOS diagnostic behavior ratings in Table 4. When using the manually segmented turn-taking events, in the tasks of "AD vs AS", "AS vs HFA", "AD vs HFA", and "AD vs AS vs HFA", our best recognition accuracies are 0.82, 0.80, 0.78, and 0.61 compared to 0.75, 0.71, 0.62, and 0.50 when using ADOS behavior codes as features, respectively. The results show a degradation in the accuracy when using the automatically-defined turn-taking events, especially in those cases where exact time boundary for durational feature computations are important. Our multimodal features computed on participant, investigator, and interaction between two can be seen as proxy measures to various granular aspects of interaction behaviors manifested in the signal characteristics, and the ADOS behavior ratings can be seen as experts perceptual judgment on participant's severity in the deficit of the social-communicative ability after being engaged in an interaction. Our recognition results indicate that the signal-based behavior measures possess as much discriminative information on "AD vs AS" and "AS vs HFA" as compared to ADOS codes; and they clearly outperforms ADOS behavior ratings on the task of differentiating "AD vs HFA".

4. Conclusions

In this work, we present a novel study using audio-video recordings of ADOS interviews. To the best of our knowledge, there has not been automatic analysis on the behavior differences among the three categorization of ASD. By computing features representing characteristics of acoustics, motions, and conversation structures of the participant and the investigator, we show that there exists significant differences between the three syndromes. In fact, our behavior features possess more discriminative power in detecting different subgroups of ASD compared to the manual behavior codes derived from the ADOS manuals. While additional works on larger sample size are needed, our initial result implicates the potential merit in deriving granular and signal-based features as quantitative measures to study socio-communicative behaviors, which are often qualitatively-described currently, at scale.

One of the immediate future work is to improve the endpoint algorithm for our automatic turn-taking events segmentation to ensure the robustness of our durational-based features. Further, since ASD is a heterogeneous condition, except for ADOS, other clinically relevant information, such as ADIR, should also be considered. Each of these instruments provides a window into better stratification of the broad ASD spectrum. We will explore the use of multi-task learning that jointly leverages multiple existing instruments along with the signal-derived behavior cues in tasks of both detecting ASD and precise behavior categorization of different ASD sub-groups.

5. References

[1] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[2] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[3] J. Baio, "Prevalence of autism spectrum disorders: Autism and developmental disabilities monitoring network, 14 sites, united states, 2008. morbidity and mortality weekly report. surveillance summaries. volume 61, number 3." *Centers for Disease Control and Prevention*, 2012.

[4] E. Marchi, B. W. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: evidence across three languages." in *INTERSPEECH*, 2015, pp. 115–119.

[5] A. Pahwa, G. Aggarwal, and A. Sharma, "A machine learning approach for identification & diagnosing features of neurodevelopmental disorders using speech and spoken sentences," in *Computing, Communication and Automation (ICCCA), 2016 International Conference on*. IEEE, 2016, pp. 377–382.

[6] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[7] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni, "Use of machine learning to identify children with autism and their motor abnormalities," *Journal of autism and developmental disorders*, vol. 45, no. 7, pp. 2146–2156, 2015.

[8] J. Diehl, K. Tang, and B. Thomas, "High-functioning autism (hfa)," in *Encyclopedia of Autism Spectrum Disorders*. Springer, 2013, pp. 1504–1507.

[9] C. Gillberg, "Asperger syndrome and high-functioning autism." *The British Journal of Psychiatry*, vol. 172, no. 3, pp. 200–209, 1998.

[10] S. Ozonoff, M. South, and J. N. Miller, "Dsm-iv-defined asperger syndrome: Cognitive, behavioral and early history differentiation from high-functioning autism," *Autism*, vol. 4, no. 1, pp. 29–46, 2000.

[11] G. B. Mesibov, V. Shea, and L. W. Adams, "Asperger syndrome/high functioning autism," *Understanding Asperger Syndrome And High Functioning Autism*, pp. 1–23, 2001.

[12] M. Ghaziuddin, "Brief report: Should the dsm v drop asperger syndrome?" *Journal of autism and developmental disorders*, vol. 40, no. 9, pp. 1146–1148, 2010.

[13] J. L. Sanders, "Qualitative or quantitative differences between asperger?s disorder and autism? historical considerations," *Journal of autism and developmental disorders*, vol. 39, no. 11, p. 1560, 2009.

[14] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[15] J. Barahona-Corrêa and C. N. Filipe, "A concise history of asperger syndrome: The short reign of a troublesome diagnosis," *Frontiers in psychology*, vol. 6, 2015.

[16] K. E. Macintosh and C. Dissanayake, "Annotation: the similarities and differences between autistic disorder and asperger's disorder: a review of the empirical evidence," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 3, pp. 421–434, 2004.

[17] K. Macintosh and C. Dissanayake, "Social skills and problem behaviours in school aged children with high-functioning autism and asperger?s disorder," *Journal of autism and developmental disorders*, vol. 36, no. 8, pp. 1065–1076, 2006.

[18] M. F. Casanova, A. El-Baz, and J. S. Suri, *Autism Imaging and Devices*. CRC Press, 2016.

[19] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[20] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "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*, vol. 30, no. 3, pp. 205–223, 2000.

[21] C. Lord, M. Rutter, and A. Le Couteur, "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*, vol. 24, no. 5, pp. 659–685, 1994.

[22] P. J. White, M. O?Reilly, W. Streusand, A. Levine, J. Sigafoos, G. Lancioni, C. Fragale, N. Pierce, and J. Aguilar, "Best practices for teaching joint attention: A systematic review of the intervention literature," *Research in Autism Spectrum Disorders*, vol. 5, no. 4, pp. 1283–1295, 2011.

[23] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013. [Online]. Available: http://hal.inria.fr/hal-00873267

[24] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 35, no. 2, pp. 205–220, 2005.

[25] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[26] M. Ghaziuddin and L. Gerstein, "Pedantic speaking style differentiates asperger syndrome from high-functioning autism," *Journal of autism and developmental disorders*, vol. 26, no. 6, pp. 585–595, 1996.