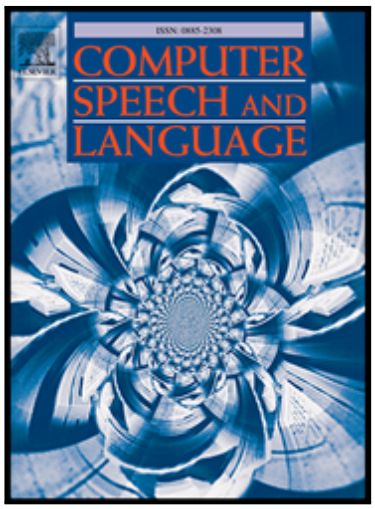

Toward Differential Diagnosis of Autism Spectrum Disorder using Multimodal Behavior Descriptors and Executive Functions

Chin-Po Chen, Susan Shur-Fen Gau, Chi-Chun Lee

PII: S0885-2308(17)30356-X
DOI: https://doi.org/10.1016/j.csl.2018.12.003
Reference: YCSLA 966

To appear in: *Computer Speech & Language*

Received date: 16 December 2017
Revised date: 14 October 2018
Accepted date: 1 December 2018

Please cite this article as: Chin-Po Chen, Susan Shur-Fen Gau, Chi-Chun Lee, Toward Differential Diagnosis of Autism Spectrum Disorder using Multimodal Behavior Descriptors and Executive Functions, *Computer Speech & Language* (2018), doi: https://doi.org/10.1016/j.csl.2018.12.003

Highlights

- Multimodal behavior descriptors computing from large-scale audio-video recordings of ADOS interviews
- Robust classification results obtained between the three types of ASD, AS, HFA, AD, using executive functions of CANTAB and multimodal behavior descriptors
- ASD subjects turn-exchange durations in spontaneous interaction are correlated with Rapid Visual Information Processing measure

Toward Differential Diagnosis of Autism Spectrum Disorder using Multimodal Behavior Descriptors and Executive Functions

Chin-Po Chen[1,3], Susan Shur-Fen Gau[2], Chi-Chun Lee[1,31]

[1]*Department of Electrical Engineering, National Tsing Hua University, Taiwan*
[2]*Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan*
[3]*MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan*

Abstract

Varied manifestations of social communication deficits, atypical prosody, and restricted and repetitive behaviors are often observed in individuals with autism spectrum disorder (ASD). The pervasiveness and heterogeneity in ASD have made it an increasingly important interdisciplinary research domain. The categorizations in ASD, ie. Autistic Disorder, High-functioning autism, Asperger' syndrome, has varied throughout the past versions of Diagnostic and Statistical Manual of Mental Disorders (DSM) in order to have a better description of ASD. Using computational approach in characterizing these neuro-developmental disorders is, therefore, important for characterizing relevant behavior constructs consistently with potential wide applicability. In this work, we propose to compute signal-derived multimodal behavior descriptors of ASD subjects during dyadic interactions of Autism Diagnostic Observation Schedule (ADOS), and we further examine these behavior features' discriminatory power in differentiating between the three groups in ASD: Autistic Disorder (AD), Asperger' syndrome (AS), and High-functioning autism (HFA). Additionally by combining assessment of ASD subject's executive functions, i.e., measured by Cambridge Neuropsychological Test Automated Battery (CANTAB), the classification accuracy improved further especially on AD versus AS. Finally, we found a moderate correlation between turn-taking duration in our computed behavior features and measures of the Rapid Visual Information Processing in CANTAB.

Keywords: behavioral signal processing, autism spectrum disorder, multimodal behaviors descriptors, executive functions, differential diagnosis

Email address: gaushufe@ntu.edu.tw, cclee@ee.nthu.edu.tw (Chin-Po Chen[1,3], Susan Shur-Fen Gau[2], Chi-Chun Lee[1,3])

URL: http://biic.ee.nthu.edu.tw/cclee.php, http://www.ntuh.gov.tw/psy/physician/shurfengau/default.aspx (Chin-Po Chen[1,3], Susan Shur-Fen Gau[2], Chi-Chun Lee[1,3])

1. Introduction

Autism spectrum disorder (ASD) is a neuro-developmental disorder, often characterized by their impaired social-communication skills with restricted and repetitive behaviors. Conducting studies into better characterizing ASD has recently gained more interest due to the symptom's increasing prevalence and its inherent heterogeneity. Reported in 2014, 1 in 68 children is being diagnosed with autism [1], and the 8-year-old children's ASD diagnosis rate has increased from 4.2% in 1996 to remarkable 15.5% in 2010 [1]. Heterogeneous impaired expressive behavior profiles have been found in ASD individuals reflecting varied manifestations of their core disabilities regarding socio-communicative deficit and restricted-repetitive interests [2, 3, 4, 5].

The initial awareness of ASD individuals' social-communicative abnormalities, often manifested behaviorally, usually starts relatively early at their infant stage (though the exact onset of autism greatly varies among individuals). Past findings indicated that these early age individuals have a delay in their language development as compared to a normal typically-developing child at the same age [6]. Social deficits are also spotted when children with autism do not come to interact with others, whereas neuro-typically developing (TD) children usually seek out for their friends or parents instead. Furthermore, repetitive behaviors are seen among ASD individuals as they tend to start repeating specific actions or focusing on local details of a picture, i.e., lines or wheels of a car, instead of the picture as a whole [2]. Aside from common clinical diagnostic criteria, such as the ICD-10 [7], the DSM-5[8], and the Gillberg and Gillberg Diagnostic Criteria [9], researchers have also developed a variety of clinically-validated instruments targeted to quantitatively assess these expressed atypical socio-communicative behaviors, mainly through two major mechanisms: self/parents report and diagnostic interviews. In specifics, the gold standard in using diagnostic interviews is the Autism Diagnostic Observation Schedule (ADOS) [10]. ADOS is a semi-structured spontaneous face-to-face interview, conducted by certified clinicians, providing a standard protocol for eliciting behaviors from the participants in order to assess their social-communicative ability. Significant group differences have been consistently demonstrated to exist between individuals with autism and TD individuals during ADOS interviews, and the ratings from the ADOS manual provide a general numerical assessment of the severity of ASD subjects' behavioral symptoms.

Being a neuro-developmental disorder, a variety of studies have also been conducted to understand the internal cognitive function of autisms, in particular, executive functions (EF). Executive functions are driven by prefrontal cortex and have been used to identify developmental disorders, especially relevant for autism [11, 12, 13, 14, 15]. Hill et al. stated that better identification on components of human's executive system by assessing a wide range of its cognitive functions is key to bring additional insights in autism [16]. One measurement of executive function is CANTAB, which is a computer-administered task set measuring visual memory, attention, and planning [17]. Two subsets in CANTAB, i.e., stockings of Cambridge (SOC) and intradimensional/extradimensional shift tasks (ID/ED), have been tested for ASD and other groups. Regarding SOC, results have shown that autism groups took significantly more moves than those of the mentally retarded and control

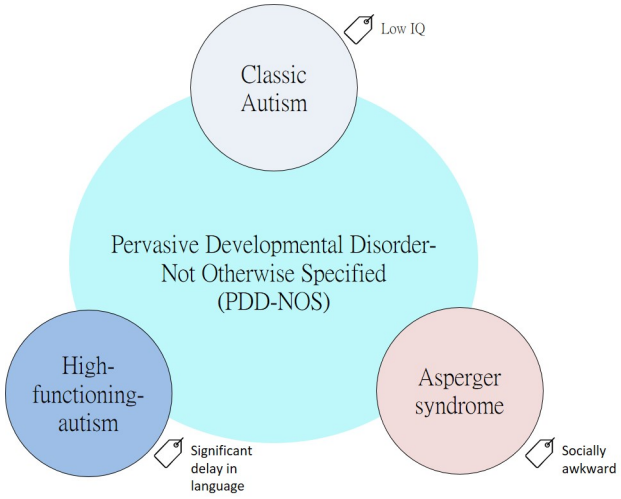

Figure 1: A brief demonstration of the relationship between classifical autism, Asperger's syndrome, and high-functioning autism. The known key differential marker between these syndroms is also shown in the figure.

groups, but the results were not significant when tested using ID/ED task [11]. Furthermore, Ozonoff et al. initially failed to detect such planning impairments when testing on the high-functioning autism group [18]. However, in a more recent study involving larger sample size, Ozonoff et al. found that measurements of SOC and ID/ED task showed significant differences between autism group and control group not in all but certain stages [11]. Steele et al. also claimed that a reduced spatial working memory in autism can be tested by Spatial Working Memory (SWM) in CANTAB [19].

While extensive research effort exists both in characterizing the social-communicative behaviors and assessing internal executive functions for ASD, its clinical definition and diagnostic criteria continue to change over time due to the heterogeneity and the complexity in characterizing ASD symptoms precisely. In fact, in DSM-4, several types of autism, including autistic disorder, Asperger's syndrome (AS), and pervasive developmental disorder not otherwise specified (PDD-NOS), which were originally specified in DSM-3, have all been grouped under an overall umbrella term, ASD [20]. Figure 1 demonstrates a general differentiation among the three clinical subgroups within ASD: classical autism (AD), high functioning autism (HFA), and Asperger syndrome (AS). Briefly speaking, AS subjects tend to have lower functioning abilities (low IQ) compared to the other two, and HFA subjects often exbihits language delay in childhood while AS subjects do not. Asperger's syndrome, nonetheless, presents awkward social behaviors compared to those of typical development[21, 22]. However, in the latest DSM-5, AS (together with PDD-NOS) has been completely eliminated [8], mainly due to the fact that the DSM-4 criteria of AS (autistic social deficit without language and cognitive delay) is non-differentiable from criteria of HFA in practice. The manifested impaired social relationship of AS and HFA has been argued to be caused by different mechanisms [23]. As an example, several studies have presented evidences that AS subjects often attempt to make interaction with other people but often

fail due to inappropriate ways of expressions [21, 22]; on the other hand, people with HFA do not show such an initative attempt to start a conversation.

Most, if not all, of the current ASD clinical assessment or diagnoses are often behaviorally-based with measurements derived from self-report (ADIR) or experts manual observation coding (ADOS). This method of *quantifying* relevant behavioral constructs often suffers from standard issues of scalability and human subjectivity [24]. With the larger-scale availability of data collection and the increasing collaboration between medical professionals and engineers, using computational methods, such as signal processing and machine learning techniques, across mental health applications have been shown to be a promising approach in transforming status quo by providing objective behavior analytics derived directly from the audio-video data [24, 25].

Furthermore, the current assessment of ASD suffers from not only subjectivity raised from the observational coding procedure but also additional complexity due to the nature of ASD behavior symptoms. For example, due to the ASD's impaired socio-communicative functions, the ADOS administrator is required to serve both as an interacting partner, i.e., to help elicit the targeted social behaviors, and also as an expert observer, i.e., to rate the severity of the impairment. While this setting has been the gold standard in clinical interviews of ASD, this particular behavior quantification method is naturally limited by its rating protocols, e.g., the behavior dynamics of the two interacting partners (the subject and the investigator) can not be explicitly measured due to the design of coding manuals and interaction procedures. Recently, a variety of research has indicated that quantifying intricate behavior dynamics between interlocutors by utilizing computational methods grounded in direct computation of behavior signals is key in bringing relevant insights beyond current measurements at scale, e.g., exemplary use case exists in applications of couple therapy [26, 27] and motivational interview therapy [28, 29].

In fact, by abstracting ADOS as a composition of two parts: 1) the social interaction protocol, i.e., the design of the various semi-structured activities in soliciting clinically-relevant behaviors through interaction, and 2) the ADOS manual observation coding, i.e., the numerical ratings of behavior constructs that the investigator needs to pay attention to during the interaction. Our past research indicated a preliminary finding that by deriving multimodal behavior descriptors characterizing both the ASD subject and the investigator behaviors directly from audio-video data collected during ADOS sessions, these signal descriptors possess substantial discriminatory power in differentiating between the three subgroups of ASD: AD vs. HFA vs. AS [30]. In this work, we extend upon our previous preliminary research in developing computational methods for differentiating between the three different groups of ASD. Specifically, our major contributions are the following:

1. Computing spontaneous multimodal socio-behavior descriptors from a larger audio-video signal database collected during ADOS interviews
2. Performing automatic categorization of the three types of ASD groups using CANTAB executive function measures and the derived multimodal socio-behavior descriptors
3. Analyzing the relationship between the derived behavior descriptors and the CANTAB executive function measures

We collect a large-scale audio-video database of 60 ASD subjects in total engaged in real ADOS interacting sessions, and our computed behavior descriptors includes multimodal aspects such as body movements, prosodic characteristics and turn-taking timing of the participants (subjects) and the investigators (clinicians), and also the dynamics between the two during the interviews. We additionally include measures of the executive functions, i.e., internal neuro-cognitive functions, on these subjects using the CANTAB to include cognitive function to complement the computed behavior descriptors in improving our automatic differentiation between the three major groups of ASD. Finally, the measures of CANTAB are analyzed with respect to the various behavior descriptors computed during ADOS interviews. To the best of our knowledge, this is one of the first systematic investigators of a large pool of ASD subjects undergoing rigorous ADOS to computationally understand and automatically differentiate the three groups of ASD using signal processing and machine learning approach with additional inclusions and analyses of measures on subjects' internal executive functions.

The rest of the paper is structured as the following: in Section 2 we will briefly introduce some past research about using social-behavior signal on research of autism, in Section 3, we will introduce our database including subjects' demographics, collection protocols, and ADOS and CANTAB descriptions. In Section 4, we will describe our research method in deriving multimodal socio-behavior descriptors from ADOS. In Section 5, we will demonstrate our experimental results and analyses, and finally, Section 6 is our conclusion.

2. Literature Review

A few notable examples of using signal processsing and machine learning techniques for the study of ASD are listed below: Bone et al. presented a computational study of spontaneous prosody during ADOS interviews demonstrating that joint modeling of interlocutors' expressive prosodic characteristics helps improve automated assessment of children's ASD severity [31, 32]. Li et al. developed automatic classification algorithms for differentiating between TD and ASD using audio features and facial expressions [33, 34]. Leclre et al. analyzed behaviors of early age autism children using automatically-derived video features in 3D dimensions during the subjects' interaction with their parents, and they demonstrated that these derived behavior descriptors are highly correlated with the CIB scores [35]. Schuller et al. published a computer-aided system that provides a platform to facilitate training of socio-emotional communication skill for autism [36]. Lastly, Ringeval et al. released a database designed to analyze speech and language characteristics of language-impaired children (LIC) and those of pervasive developmental disorder [37], and this database has further been used in the past INTERSPEECH challenge [38, 39].

3. Database Description

Our database used in this paper includes two different instruments in assessing ASD subjects, i.e., ADOS and CANTAB. We will briefly describe each of them and the collection procedure in the following section.

Table 1: (Left) A detailed list of ADOS activities in Module 3 and Module 4 ('*' means optional). (Right) A detailed list of all assessments in the CANTAB[17]. There are warm-up tasks: {MOT, BLC}, visual memory: {DMS, PAL, PRM, SRM}, execution function, working memory, planning: {AST, IED, OTS, SSP, SWM, SOC}, attention: {CRT, MTS, RVP, RTI, SRT}, decision making and response control: {CGT, IST, SST, ERT}

ADOS		CANTAB
Module 3	Construction Task, Make-believe Play, Demonstration task, Description of a Picture, Telling a Story From a Book, Cartoons, Conversation and Reporting, Emotions, Social Difficulties and Annoyance, Break, Friends and Marriage, Loneliness, Creating a Story	Motor Screening (MOT), Big/Little Circle (BLC), Delayed Matching to Sample (DMS),Paired Associates Learning (PAL), Pattern Recognition Memory (PRM), Spatial Recognition Memory (SRM), Attention Switching Task (AST), Intra-Extra Dimensional Set Shift (IED), One Touch Stockings of Cambridge (OTS), Spatial Span (SSP), Spatial Working, Memory (SWM), Stockings of Cambridge (SOC), Choice Reaction Time (CRT), Match to Sample Visual Search (MTS), Rapid Visual Information Processing (RVP), Reaction Time (RTI), Simple Reaction Time (SRT), Cambridge Gambling Task (CGT), Information Sampling Task (IST), Stop Signal Task (SST), Emotion Recognition Task (ERT)
Module 4	Construction Task*, Telling a Story From a Book, Description of a Picture*, Conversation and Reporting, Current Work or School*, Social Difficulties and Annoyance, Emotions, Demonstration task, Cartoons*, Break, Daily Living*, Friends and Marriage,	

3.1. Autism Diagnostic Observation Schedule (ADOS)

ADOS is a gold standard for assessing the severity of autism using the observational approach in a semi-structured face-to-face interview session. There are four different modules (M1 to M4) of ADOS, where each module is designed for subjects with different language developmental levels. All of our subjects fit the criterion to be eligible to participate in either M3 or M4. ADOS is designed with 14 different tasks; for example, these tasks involve telling a story from a picture book, spoken interactions about emotional experiences, a demonstration task, etc. (a complete list can be found in Table 1). The setting of this diagnostic interview is that two people, an investigator (trained psychologists/clinicians) and a participant (ASD subjects, most of them are teenagers), involve in an face-to-face interaction. The investigator is asked to both rate the participant's behaviors and be in the role to interact with the participant.

Table 2 shows the contents of the ADOS coding under four categories: language and communication, reciprocal social interaction, play + imagination/creativity, stereotyped behaviors and restricted interests. Language and communication measures the participant's ability to convey to the investigator such as pointing (PNT), reporting events (REPT). Reciprocal social interaction measures how the participants behave when they receive attention from the interacting partner, and codings like spontaneous initiation of joint attention (IJA) is used; play + imagination/creativity measures the participant's imagination when telling a picture story of their creation. Stereotyped behaviors and restricted interests are rated by observing their verbal and non-verbal responses to during the question-answering interactions. The whole ADOS process lasts forty minutes to one hour. ADOS can be conceptualized as two major components, i.e., a social interaction protocol (to elicit behaviors from the subjects through interaction) and manual behavior rating (to numerically assess behaviors of the subjects with manual rating).

Social interaction. We focus on the 'emotion' session of ADOS, which involves mainly spoken interactions. During the emotion session, the participants are asked to describe and share episodes of their experienced feelings such as angry and happy. An emotion session lasts

about two to ten minutes depending on how the participant finishes answering the set of questions posed by the investigator. The situation is a back-and-forth dialog between the investigator and the participant, and most of the conversations are initiated by the investigator.

Behavior rating. During ADOS sessions, the inspector conducts the designed social activities and writes down his/her observations about the participants' behaviors with respect to different items listed in the ADOS manuals. The behavior ratings are based on 28 items assessing the communication skills, social interaction skills, and the restricted/repetitive interest. Then, a final communication score and a social score is computed from these itemized behavior ratings as an overall assessment of the communication ability and social ability. These manually-coded behaviors can be thought as a measurement of the investigators' observations on the participant's behaviors.

Audio-Video Data Collection. We collect audio-video recordings of ADOS sessions at the National Taiwan University Children Hospital [4]. We set up three high-definition cameras (one facing the participant, the other one facing the investigator, and the third one capturing the two people from the side), and two lapel microphones (each clipped on the collar of an individual speaker). Figure 2 shows a mock-up scene of our ADOS data collection from two different camera views. The two audio channels from each microphone are synchronized through an audio processing mixer, and we also synchronize the video with audio manually with a clap board. In total, we collect 60 sessions of ADOS. The clinical diagnoses for each of the subjects, i.e., either classical autism (AD), Asperger's syndrome (AS), or high-functioning autism (HFA), is determined using a combination of diagnostic tools (e.g., ADIR, ADOS, other in-clinic interactions). The utterances in the ADOS are manually segmented. Table 3 shows the demographics of our ADOS subjects.

3.2. Clinical measurement of executive function – CANTAB

CANTAB is a computerized analytic tool used in testing participant's executive function, such as working memory and sustained attention [17]. It has been used on a variety of neuro-developmental disorders: ASD, Attention Deficit Hyperactivity Disorder (ADHD),

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

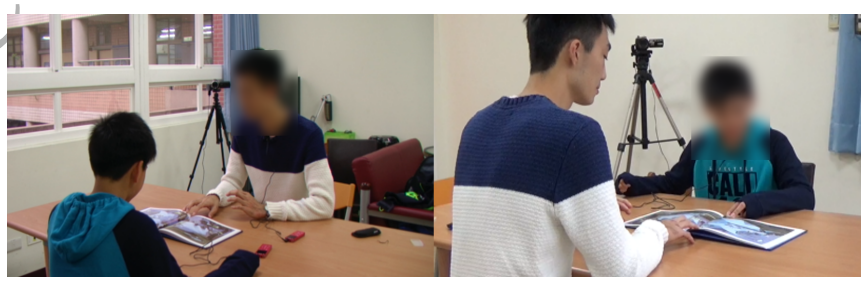

Figure 2: A mock view of our ADOS audio-video data collection setting

Table 3: Demographics of our ASD participants in our dataset: the value in the parenthesis indicates the number of subjects

	Subjects Demographics		
Age (Avg/Std)	Autism	AS	HFA
ADOS ($n = 60$)	$15.03 + / - 3.08(28)$	$15.95 + / - 3.2(20)$	18.5+/-4.4 (12)
CANTAB ($n = 52$)	14.77+/-3.23(21)	15.55+/-3.21(20)	19.36+/-2.83(11)

and mental retardation [40, 41, 42]. The CANTAB testing items are listed in Table 1, among the testing items, some of them are categorized. MOT and BLC are simple training tasks for latter tasks. DMS, PAL, PRM, SRM are categorized as visual memory tasks; AST, IED, OTS, SSP, SWM, SOC are categorized as measurement for execution function, working memory, and planning tasks; CRT, MTS, RVP, RTI, SRT are categorized as attention tasks. Finally, CGT, IST, SST, ERT are categorized as decision making and response control tasks. Items of the CANTAB analysis are listed in Table 1. The total number of subjects who have gone through both ADOS and CANTAB is 52 (less than the total 60 ADOS subjects). We also list this distribution in Table 3.

4. Research Methodology

In this section, we will describe our multimodal behavior descriptors extraction approach applied on the ADOS audio-video recordings. Figure 3 shows a systematic diagram. The complete procedure involves the following: low-level audio and video feature extraction (LLDs), segment-level feature encoding on LLDs with respect to the investigator, the participant, and the inter-personal dynamics (turn-taking duration measures), and finally we derive a session-level (i.e., a vector representation of behavior characterization over a complete emotion session) that is used for analyses and automatic ASD group categorization. An example of one session-level feature is denoted as:

$$\sigma - [Pitch_{invest}^{investquest}] \tag{1}$$

It means we compute the pitch LLD, encode it at the segment-level (discussed in section 3.2) using standard deviation (σ). The superscript: investquest, denotes the region where the investigator is speaking. The subscript: 'invest' means that this kind of feature is calculated with respect to the investigator.

4.1. Audio-Video Low-Level Descriptors (LLDs)

The low-level descriptors are extracted from both audio and video data. We compute normalized body action energy (NBAE) as our motion LLDs from the video, prosodic characteristics as acoustic LLDs from the audio data, and further turn-taking duration as interconversation dynamics features.

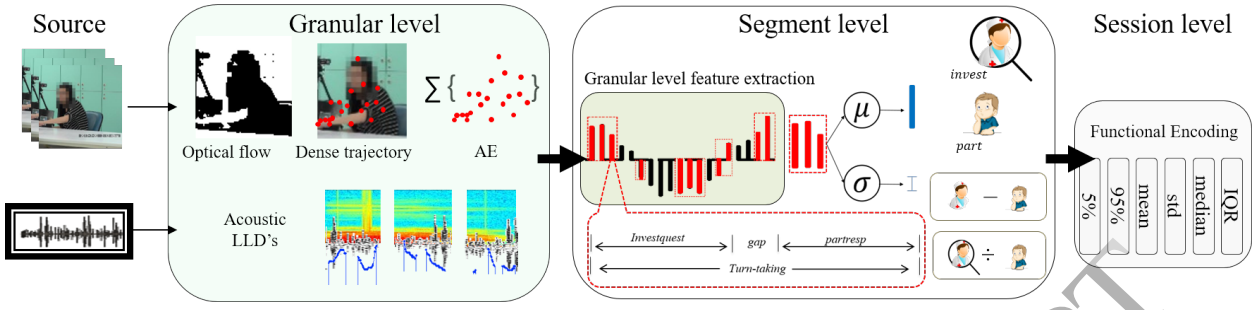

Figure 3: A systematic view of our social-behavior descriptors extraction

Motion LLDs. We compute normalized body action energy (NBAE) to measure the amount of a person's movement at the frame-level. NBAE is computed using the following steps. First, we extract trajectory points (x_t, y_t) by applying median filtering kernel M on the vertical and horizontal component of dense optical field $\omega = (u_t, v_t)$.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(x_{\bar{t}}, y_{\bar{t}})} \tag{2}$$

P_{t+1} where the dense sampling is implemented with the method of dense trajectory tracking described in [43]. This method has been shown to be successful in deriving descriptors for action recognition [44, 45].

By summing up these tracked trajectories, it can be thought of as a representation of the amount of movement. We first compute $P(i)$ as the summation of the total number of moving trajectories for 15 frames (≈ 0.5 seconds), termed as the *action energy* (AE) in this paper. Then, the AE is normalized with respect each person (session) to derive the NBAE,

$$NBAE(i) = \frac{\overline{P(i)} - \mu}{\sigma} \tag{3}$$

where μ and σ are computed for that particular individual over the session. This descriptor is computed at a 0.5 second increment. NBAE is a one-dimensional feature, which can be thought as the relative amount of movements with respect to an individual's baseline.

Acoustic features. Regarding acoustic features, we compute low-level prosodic descriptors (LLDs), including pitch, intensity, the harmonic-to-noise ratio (HNR), jitter, and shimmer (i.e., a five-dimensional vector per frame) using the Praat toolkit [46]. All of the acoustic features are extracted every 10ms; these LLDs are further z-normalized with respect to an individual speaker.

1. Frequency related features: Pitch, Intensity (loudness)
2. Voice quality related features: Jitter, Shimmer, HNR

This set of LLDs has been used to characterize paralinguistic acoustic parameters for a variety of automated recognition tasks. For example, pitch contour and energy can serve as an effective indicator for vocal emotion states [47]. Computing voice quality to characterize

damaged voices like breathiness and harshness have been shown to be measured by HNR [48, 49], and also jitter and shimmer [50]. Furthermore, children with ASD have also been reported to exhibit atypical acoustic characteristics: extreme ranges of intensity and pitch level, considerable hoarseness harshness hypernasal sound [48, 49]. Researchers have used low-level prosodic descriptors with statistical functionals computed over a duration in order to capture these atypical prosodic characteristics [50]. Recently, Bone et al. have further demonstrated that both the acoustic parameters of ASD subjects and the acoustic parameters of the investigator are indicative of the subject's ASD severity level during the 'emotion' part of the ADOS interviews [32].

Turn-taking durational features. When people engage in conversations, speakers exchange coordinated turn-taking to talk non-simultaneously [51]. Past research has shown that turn-taking deficits, e.g., awkward pause or inappropriate use of turn-taking cues, exists in autistic people [52]. Intervention has been developed to educate children with ASD to engage in appropriate turn taking in conversations [53, 54].

Turn-taking regions in this research are defined at each turn exchange during the emotion session of the ADOS interviews. Most of the spoken interaction in our data start with the following situation: first, the investigator initiates a question and the participant will respond. A *turn exchange*, termed as a turn-taking region, is defined from the start of investigator's question to the end of the participant's response. We split turn-taking region into three parts, Investquest, Gap, Partresp. Refer to Figure 4, Investquest is defined in the region that the investigator initiates a question to the end of the question. Gap is defined from the end of investigator's question to the beginning of the participant's response. Finally, Partresp is defined in the region where the participant's response to the investigator's question. Then we compute turn-taking durational feature as the time duration within each of the specified segments. There are situations where the participant's response precedes the end of Investquest causing speech overlap. In this situation, the durational feature computed over the Gap region will be negative.

4.2. Segment-Level Features

We encode the low-level descriptors mentioned above to the segment-level features using mean and standard deviation, where the segment is defined as each of the regions in a turn-taking exchange. Within a turn-taking, we derive features for the investigator and the participant as additional measures on inter-personal dynamics. This segment-level feature extraction approach will result in three different perspectives: Intra-Invest (investigator speaking), Intra-Part (participant speaking), Inter-Behavior (differences between the two

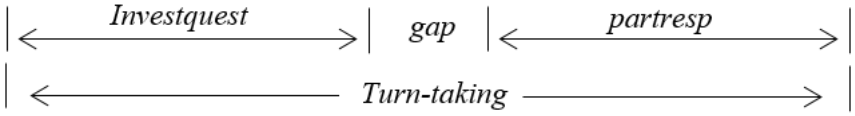

Figure 4: A brief description of turn-taking segments defined in this research

Table 4: A list of different basic operators that are used to derive the multimodal behavior descriptors

Session-level Functional	LLD Features	Segment-level Functional	Region of Computation	Perspectives
5% percentile	NBAE	mean (μ)	Invest_quest	Intra-Invest
95% percentile	Pitch	std (σ)	Part_resp	Intra-Part
mean	Intensity		Gap	Inter-behavior
median	HNR		Turn-taking	
std	Jitter			
IQR	Shimmer			
	Duration			

Intra-features within a turn region). We will describe segmental features for each behavior modality separately below.

Segmental NBAE. We compute the mean value of frame-based NBAE in each turn-segment. We then further derive the inter-relationship of NBAE between the two interlocutors at the unit of a segment. For example: μ-[$\text{NBAE}_{inter}^{investquest}$] represents segmental NBAE calculated in the region: Investquest (denoted by the superscript), and the subscript: 'inter' means it calculates the inter-relationship of investigator's and participant's feature. The inter-relationship is computed by subtracting investigator's NBAE from the participant's one. The total number of segmental NBAE features is 12.

Segmental acoustics features. For acoustic modality, we compute mean and standard deviation in each of the turn-segment perspectives for each LLD. Similar to NBAE, we also calculate the inter-relationship of acoustic features between the interlocutors. We first average the low-level acoustic descriptors, i.e., pitch, intensity, HNR, over a 0.5-second window (in order to synchronized framerate with NBAE), then, we compute mean and standard deviation as the segmental function encoding in each segment. As an example, μ-[$\text{Intensity}_{invest}^{investquest}$] means the intensity of investigator's acoustics averaged within Invest_{quest}. The inter-relationship of acoustic LLD's are also calculated within a segment, for example: μ-[$\text{Pitch}_{inter}^{investquest/partresp}$] means the division of averaged pitch of Invest_{quest} over Part_{resp}. The total dimension of segmental acoustic features is 12.

Segmental turn-taking features. We take the duration of Investquest, Gap, Partresp and the entire turn-taking as our basic turn-taking features, and we further compute the inter-relationship (ratio) between each segments. For example: Investquest / Gap, Investquest / Partresp. This results in a total of 9-dimensional features representing turn-taking characteristics at the segmental level.

4.3. Session-Level Features

Finally, segment-level features are further encoded to session-level features using a variety of robust functionals: 5% percentile, 95% percentile, mean, median, standard deviation (std), IQR, in order to describe the distribution of the derived segmental multimodal behavior features. A similar approach has been developed in the past to perform acoustic-based emotion analysis [55]. These are the features used in the final differential classification experiments in categorizing between the three groups of ASD. A list of different parameters

used in deriving our final session-level features (from frame-level, segment-level, to session-level) can be found in Table 4.

5. Experimental Setup and Results

We conducted the following three different experiments in this work.

- **Experiment I:** Classification between the three types of ASD clinical diagnosis (AD, AS, HFA) using multimodal behavior features computed from the ADOS recordings
- **Experiment II:** Classification between the three types of ASD by fusing behavior features with subsets of the CANTAB measures
- **Experiment III:** Correlation analyses between ASD subject's multimodal behavior features computed from the ADOS and the executive function measures derived from the CANTAB

The classifier used is logistic regression in this work, and we further report results using support vector machine (linear kernel) and random forest for each experiment in Table 5, 6, 7, and 8. Logistic regression is used to avoid potential issue of overfitting due to the sample size for each class of ASD subgroup. In addition, we carried out stepwise regression to select the most relevent features in our recognition tasks, which is based on method of univariate feature selection that calculates F-value to determine the importance of each feature. Table 6 (bottom) shows the classification accuracy of the selected CANTAB features within three tasks. The evaluation scheme is leave-one-subject-out cross-validation, and the metric used is the unweighted average recall (UAR).

5.1. Experiment I Results and Discussions

We use the session-level multimodal behavior features to perform our classification task. Similar to our previous study [30], our baseline is to compare these signal-derived features to manual observational behavior scores derived from the ADOS manual and executive function measurement of CANTAB.

A summary of the multimodal classification results is provided in Table 9. The best classification performance is highlighted in red, which is 0.68, 0.80, 0.76, and 0.54 for AD vs. AS, AS vs. HFA, AD vs. HFA, and AD vs. AS vs. HFA respectively. These signal-derived behavior descriptors outperforms ADOS behavior ratings (communication score and social reciprocity score), which achieves only 0.65, 0.46, 0.60, and 0.43. The use of multimodal behavior features outperform single behavior modality (results listed in Table 5 (left)). The best multimodal behavior feature set for AD vs. AS task is μ-[NBAE$^{gap}_{inter}$] +μ-[Intensity$^{iq}_{i}$] +Duration$^{iq}_{intra}$. Two of the three types of features are from the investigator's question region, i.e., at the region of investigator's speaking. μ-[NBAE$^{pr}_{i}$] +μ-[Intensity$^{iq}_{i}$] +Duration$^{gap/iq}_{inter}$ and μ-[NBAE$^{gap}_{inter}$] +μ-[Intensity$^{iq}_{i}$] +Duration$^{gap}_{intra}$ also shows good recognition accuracy in this task. Since most of the features emerged from investquest region, this might imply the investigator's behavior reflects the difference between AD and AS.

Table 9: Multimodal classification on the designed tasks. Bolded value means its accuracy value is higher than baseline (ADOS Communication, Social Reciprocity), and the highest value in each task is highlighted in red color. The meanings of abbreviations are listed below, AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism

F-(action,acoustic, turn-taking)	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$_i^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.47	**0.74**	0.65	0.45
μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$	0.54	0.51	0.76	0.43
μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	**0.66**	0.58	0.57	0.54
μ-[NBAE$_i^{pr}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.49	**0.75**	0.71	0.42
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.53	0.55	0.67	**0.52**
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{intra}^{tt}$	0.49	0.60	**0.73**	0.43
μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$	0.39	0.80	0.65	0.40
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{iq}$	0.68	0.60	0.38	0.34
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{gap}$	**0.66**	0.58	0.59	0.41
μ-[NBAE$_{inter}^{gap}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{pr}$	0.47	**0.75**	0.60	0.39
μ-[NBAE$_{inter}^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.50	**0.71**	0.60	0.37
eGeMAPS [57]	0.58	0.67	0.57	0.48
ADOS (Communication, Social Reciprocity)	0.65	0.46	0.60	0.43

The best multimodal behavior feature set for task is AS vs. HFA, μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$. By inspecting other bolded values (bolded value indicates accuracy higher than the baseline), we found that σ-[Pitch$_p^{pr}$] plays an important role in discriminating between AS and HFA. The descriptor represents participant's diversity of intonation (σ-[Pitch]). These could have been attributed to the observation that AS participant may possess better social skills in holding a smoother back-and-forth question-answering spoken interactions [56]. In addition, the UAR of AS vs. HFA prediction using σ-[Pitch$_p^{pr}$] alone is 0.63 (refer to Table 5). μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$ achieves the best recognition rates for the task of AD vs. HFA. The aperiodicity of μ-[HNR$_p^{pr}$] represents the impaired voice quality that could be caused by hoarse and harsh sound in speech, this feature alone achieves a classification accuracy of 0.61 (refer to Table 5), and the accuracy improves when fusing with other two modalities. Finally, μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$ achieves the best recognition rate (UAR of 0.54) for the task of AD vs. AS vs. HFA. Together with μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$. Interestingly, we found that most of the features are calculated among investigator's question region (iq) or inter-relationship that 'iq' region is involved in. This result reinforces finding in Bone et al. [32], where they showed that the investigator's prosodic features can be even more indicative of ASD subject's severity than the subject's features themselves during ADOS interviews.

Analysis. We further conduct t-test on the behavior features listed in Table 5 ($\alpha \leq 0.05$) between each pair of groups. Table 10 provides a summary on the behavior difference among these different subgroups of ASD. In task AD vs. AS, the investigator has a larger NBAE value when interacting with AS than AD during the Gap region. This descriptor represents the amount of relative movement. Therefore, it reflects the relative movement of the investigator is larger when interacting with AS than with AD, when waiting for the

Table 10: Table of significant difference features and the directions. We subtract the first group to the second group, where the first and the second group are denoted in the superscript of the task. The meanings of abbreviations are listed below: AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism

Motion	Descriptor	AD[1] vs AS[2]	AS[1] vs HFA[2]	AD[1] vs HFA[2]
μ-[NBAE$^{gap}_{invest}$]	point max	-0.04	0.47	0.43 *
	point mean	-0.05 *	0.06	0.01
	point median	-0.1	0.26 *	0.16
	point std	-0.11	0.16 *	0.05
μ-[NBAE$^{gap}_{inter}$]	point median	0.19	-0.29 *	-0.1
	point min	0.09	-0.43 *	-0.34 *
Voice quality	Descriptor	AD[1] vs AS[2]	AS[1] vs HFA[2]	AD[1] vs HFA[2]
μ-[HNR$^{partresp}_{part}$]	point IQR	-7.46	-7.61	-15.07 *
	slope IQR	-0.0	-0.0	-0.01 *
Intonnation	Descriptor	AD[1] vs AS[2]	AS[1] vs HFA[2]	AD[1] vs HFA[2]
σ-[Pitch$^{partresp}_{part}$]	curvature IQR	-0.03	-0.31 *	-0.34 **
	curvature max	0.27	-0.84 **	-0.57
	point IQR	-0.09	-0.14	-0.23 **
	slope IQR	-0.1	-0.25	-0.35 **
	slope max	0.31	-0.85 **	-0.55
σ-[Intensity$^{investquest}_{invest}$]	curvature IQR	0.06*	-0.03	0.02
	point IQR	0.02	0.03	0.05 *
	point max	0.08	0.08	0.15 *
	slope IQR	0.06 *	-0.01	0.05
	slope std	0.04 *	-0.01	0.03
Turn taking time structure	Descriptor	AD[1] vs AS[2]	AS[1] vs HFA[2]	AD[1] vs HFA[2]
Duration$^{gap/investquest}_{inter}$	point median	-0.05	-0.24 *	-0.29 **
Duration$^{gap}_{intra}$	point median	-0.28	-0.6	-0.88 *

subject response (Gap region)(μ-[NBAE$^{gap}_{invest}$]). Meanwhile, in terms of acoustic properties, the std of voice intensity (loudness) is higher in group AD. This implies that investigator's voice variation is higher in AD than in group AS. In task AS vs. HFA, analysis in motion feature implies that the median of NBAE (μ-[NBAE$^{partresp}_{invest}$]) is higher in AS than in HFA, and μ-[NBAE$^{gap}_{inter}$] has the opposite result. This indicates that higher μ-[NBAE$^{partresp}_{invest}$] is the deterministic factor showing the difference of AS and HFA, and it represents that the investigator shows a relative more movement during the portion of participant's response. Together with the result of task AD vs. AS, we can infer from the result that the investigator have more movement when interacting with participants in AS group.

Furthermore, analysis of intonation shows that maximum slope and curvature of participant's pitch together with IQR of curvature are higher in HFA. Analysis of turn-taking time structure suggests that Duration$^{gap/investquest}_{inter}$ shows lower value in AS. This could result from two reasons, either shorter duration in Gap or longer investigator's speaking duration. Both of the conditions suggest the AS participant has a higher tendency in speaking more and engage in a more 'interactive' dialogs. In task of AD vs. HFA, the maximum of μ-[NBAE$^{partresp}_{invest}$] remains higher in AD. Voice quality analysis shows a higher variation of participants voice quality in group HFA. Analysis of pitch also shows a similar result, indicating that HFA has

Table 6: Left: Classification UAR using multimodal behavior on the 52 available subjects. Right: Classification UAR after fusing behavior features with CANTAB measures. The meanings of abbreviations are listed below μ: mean, σ: standard deviation, AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism

F-(action,acoustic, turn-taking)	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$_i^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.51/0.78	0.69/0.88	0.65/0.74	0.46/0.49
μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$	0.61/0.66	0.48/0.69	0.68/0.70	0.36/0.46
μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.73/0.83	0.56/0.83	0.58/0.72	0.48/0.60
μ-[NBAE$_i^{pr}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.51/0.78	0.74/0.86	0.61/0.77	0.38/0.49
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.51/0.78	0.55/0.78	0.58/0.72	0.41/0.52
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{intra}^{tt}$	0.39/0.68	0.62/0.76	0.58/0.70	0.31/0.34
μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$	0.39/0.73	0.74/0.86	0.65/0.74	0.36/0.46
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{iq}$	0.56/0.76	0.58/0.85	0.35/0.72	0.29/0.50
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{gap}$	0.63/0.83	0.53/0.81	0.61/0.72	0.39/0.51
μ-[NBAE$_{inter}^{gap}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{pr}$	0.39/0.76	0.74/0.78	0.58/0.70	0.30/0.42
μ-[NBAE$_{inter}^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.49/0.83	0.69/0.85	0.65/0.72	0.34/0.49
Stepwise(CANTAB)	0.76	0.81	0.74	0.54

a higher variation in vocal characteristics from the beginning to the end of the interview. The intonation of the investigator, however, shows higher value in AD than in HFA. This might suggests that the investigator's loudness is higher when interacting with subjects in AD over the entire ADOS interview. Finally, analysis of turn-taking time structure suggest that duration of Gap and ratio of Gap divided by investquest is both higher in HFA group.

5.2. Experiment II Results and Discussions

Since there are fewer data samples in CANTAB than in the audio-video ADOS data, the results of the fusion shown in Table 6 are for 52 subjects only. The result to the left of slash line in Table 6 shows the UAR obtained by using multimodal behavior features only on the 52 subjects, and the result to the right is the UAR score after fusing with CANTAB features using feature concatenation. Simply using measures of CANTAB obtains a classification accuracy of of 0.76, 0.81, 0.74, and 0.54 for AD vs. AS, AS vs. HFA, AD vs. HFA, and AD vs. AS vs. HFA, respectively. The stepwise regression result demonstrates that subsets of features in Paired Associates Learning (PAL), Pattern RecognitionMemory (PRM), Spatial Span (SSP), and Rapid Visual Information Processing (RVP), are important in obtaining good prediction accuracy for AD vs. AS; subsets of Delayed Matching to Sample(DMS), PRM, Stockings of Cambridge (SOC) are good in predicting task AS vs. HFA. Subsets of PAL, DMS, SSP, Spatial Working Memory (SWM), RVP, Motor Screening (MOT) are better at classifying between AD vs. HFA. Finally, PAL, DMS, SOC, SWM are good for the task of AD vs. AS vs. HFA.

The overall accuracy improves by fusing behavioral descriptors with subsets of CANTAB data. Some of the results are even better comparing to using subsets of CANTAB data alone. Suggesting from the result, we suppose that behavioral descriptors have complementary and correlated information with CANTAB, and thus we will show the correlation between the two different types of descriptors in Experiment III.

5.3. Experiment III Results

We further analyze the correlation between ASD subjects' multimodal behavior features (expressed, recorded, and computed from ADOS audio-video recordings) and their internal executive function measures (CANTAB). Table 7 shows the Pearson's correlations computed between signal-derived multimodal behavioral features and measures of CANTAB. We only report correlations over 0.50.

Correlation between the interaction-based behavior feature of $\text{Duration}_{inter}^{gap/investquest}$ and the two Rapid Visual Information Processing (RVP) measurements are observed in our database. $\text{Duration}_{inter}^{gap/investquest}$ is positively correlated with RVPfaP and negatively correlated with RVPB. The RVP tests are often used to measure cognitive ability in working memory and sustained attention [58, 59]. Past researches have shown impairments in sustained memory and working impairment for ASD population. For example, Ozonoff et al. examine working memory in samples of high-functioning autism [60], and Hellen et al. suggest that sustained attention of autistic disorder may come from the reluctance of dealing with externally imposed tasks [61]. A higher value in the feature $\text{Duration}_{inter}^{gap/investquest}$ corresponds to shorter duration in the investigator's question and/or longer Gap time when the participant responds to their investigator. These two factors usually present the conversation to be conducted with only questions and answer and lesser chatting. In our dataset, we show that this feature is higher in HFA than in AD and AS. A hypothesis for this phenomenon is that the HFA participant is capable of carrying out smooth conversation but no intention of having other topics of conversation. In a nutshell, this differential internal cognitive impairment in ASD subjects may have also been manifested in the behavioral measures of the turn-taking durational features, resulting in a seemingly interrogative conversational turn-taking. However, a full detailed study will be needed to further understand the relationship between impaired executive function and its manifestation in the behaviors during social interaction for ASD subjects.

The measurement of Delayed Matching to Sample (DMS) has negative correlation (with $p < .001$) to feature μ-[$\text{HNR}_{invest}^{investquest}$], σ-[$\text{HNR}_{invest}^{investquest}$] (calculated on the investigator's speech). To be more specific, the percentage and total correct rate of delayed match sample have the opposite correlation with the slope of the measurement on investigator's HNR values among each turn region. In addition, Delayed Matching to Sample (DMS) test is designed for testing visual memory and is related to attention function [62]. Participant's attention might play an important role such that when a participant with low attention function will cause an investigator to frequently vary her acoustic behavior during interaction as manifested in the voice quality measures. Finally, the measurement of 'speed of movement' SOCstT2 tested under Stockings of Cambridge (SOC) test, shows a negative correlation (with $p < .001$) to behavior feature of σ-[$\text{Pitch}_{invest}^{investquest}$] (calculated on the investigator's speech). This acoustic descriptor represents the variation level of local pitch (slope) over the entire emotion session. On the other hand, SOC is a task that depends on working memory [63]. In consequence, the result implies that lesser pitch variation in turn segments corresponds to better function of working memory. Perhaps the investigator might be able to ask straightforward questions for the participant due to the better executive function

that the participant possess on focusing on the retrieval question (e.g., recalling on personal emotion episodes in the past) that they are supposed to describe to the investigator.

6. Discussion

As observed from our statistical analyses, AD subjects differential behaviors can be better observed from investigator's variation of of loudness (σ-[Intensity$_{invest}^{investquest}$]). HFA is also an autistic disorder in subjects with higher executive function, and the above observation coincides with the fact that AS and HFA have higher level of cognitive functioning compared to AD (no delayed in language and any cognitive development) [18]. AS participant can also be distinguished by observing the investigator's amount of movement (especially at the Gap region). AS participants are reported to have inappropriate ways to interact with people [21, 22] but still maintain normal cognitive functioning level [18] as compared to AD subjects, and AS subject is also known to have higher intention to engage in social interaction as compared with HFA [22]. Finally, features related to pitch variation (σ-[Pitch$_{part}^{partresp}$]), and voice harmonicity (μ-[HNR$_{part}^{partresp}$]) in HFA participant are higher than the other two ASD subgroups. This might be related to execution function affected by attention and working memory [41], but a more detailed examination should be conducted to confirm this initial exploration. Lastly, the ratio of time duration computed between the 'Gap' and the 'investquest' region is significantly higher in HFA group, smoother turn-taking skill is assumed to be correlated with lower of this value. However, AS group still demonstrates a relatively lower value. A more detailed investigation is needed to explain this observation.

In summary, we observe that our multimodal behavior descriptors computed from the ADOS recording indeed possess significant discriminatory power in differentiating between the three different diagnoses within ASD [(Manifestated by t-test) and are able to perform well prediction results together with logistic regression classifier](i.e., pair-wise classification of 0.68, 0.80, and 0.76 with three-way classification UAR of 0.54). We further observe that not only is the behavior of the ASD subjects important, but also their interviewers (i.e., the investigators) and even the dynamics between the two are important in differentiating between these diagnoses. Because the expressive behavioral nuances between the three ASD groups can be subtle, the use of signal-derived behavior measures may potentially be a more powerful approach in capturing such a difference. Furthermore, Experiment II shows that the combination of expressed behavioral features and internal executive functional descriptors help improve the three subgroups of ASD categorization accuracy, suggesting that the designed signal derived features provide additional information to the existing clinical-relevant cognitive function testing instrument (CANTAB). In the Experiment III, on the other hand, we demonstrate that the internal deficit of cognitive function is correlated with the exhibited multimodal behaviors during ADOS clinical interviews.

7. Conclusions and Future Works

The heterogeneous symptoms in the ASD population have consistently been a key issue in proper clinical stratification for targeted intervention. In fact, despite past researchers

Table 7: A list of behavior features showing significant correlation to measures of CANTAB (*p<.05 **p<.01 and the correlation is measured using Pearson's correlation). Only correlations higher than 0.5 are listed.

Signal-derived Behavior Features	Descriptor	CANTAB	Correlation
$\text{Duration}_{inter}^{gap/investquest}$	point min	RVP: RVPB	-0.61***
$\text{Duration}_{inter}^{gap/investquest}$	point min	RVP: RVPfaP	0.60***
μ-[$\text{HNR}_{invest}^{investquest}$]	curvature median	RVP: RVPfaP	-0.53***
μ-[$\text{HNR}_{invest}^{investquest}$]	slope median	DMS: DMSpcS	-0.54***
μ-[$\text{HNR}_{invest}^{investquest}$]	slope median	DMS: DMStCS	-0.55***
σ-[$\text{HNR}_{invest}^{investquest}$]	curvature median	RVP: RVPfaP	-0.56***
σ-[$\text{HNR}_{invest}^{investquest}$]	slope median	DMS: DMSpcS	-0.57***
σ-[$\text{HNR}_{invest}^{investquest}$]	slope median	DMS: DMStCS	-0.58***
σ-[$\text{Pitch}_{invest}^{investquest}$]	curvature max	SOC: SOCstT2	-0.59***
σ-[$\text{Pitch}_{invest}^{investquest}$]	slope max	SOC SOCstT2	-0.60***

Note. Positive correlations means that increasing descriptor goes with increasing CANTAB value.
***p<.001

have demonstrated several differences between AD, AS, and HFA at various development stages, a new version of diagnosis tool, DSM-5, have re-defined the criteria. The New criteria makes AS and HFA no longer be differentiably-identified in order to make the clinical assessments consistent. In this work, we propose to differentiate the three different categorizations of ASD groups by computing spontaneous multimodal behavior descriptors computed from the real ADOS recordings directly with measures on executive function derived from the computerized task of CANTAB. The signal-based multimodal behavior descriptors characterize the participant, the investigator, and the joint behavior dynamics beyond what is explicitly captured in the coding manuals. In fact, our signal-derived features include body movements, prosodic characteristics, and turn-taking durational statistics. Our experiments show that a promising accuracy can be achieved (0.68, 0.8, 0.76 and 0.54) in tasks of AD vs. AS, AS vs. HFA, AD vs. HFA, and AD vs. AS vs. HFA, respectively, well above using the behavior rating derived from the ADOS coding.

Furthermore, executive function measured derived from CANTAB also help further differentiate between the three subgroups since subjects with AS or HFA possess higher cognitive function. By fusing multimodal behavior features with CANTAB measure, the accuracy of AD vs. AS can be improved to be 0.83. We also show that measures on Rapid Visual Information Processing are correlated to the durational statistics during turn-taking when an ASD subject is engaged in spontaneous spoken interactions of ADOS; Delayed Matching to Sample correlates to the level of voice quality variation measured on investigator; Stockings of Cambridge correlates to variation level of local pitch. This preliminary study is one of the first works in systematically computing behavioral signals from a large scale ADOS collections of audio-video data and fusing with executive function measures toward differential diagnoses between the three groups of ASD. There are multiple future directions. One

of the continuing work is to collect more ASD subjects with a wide range of other clinical instruments, e.g., ADIR, to bring a fuller picture on the symptoms of ASD, also typically developing(TD) controls will be collected in order to realize these computational methods in real clinical diagnoses settings in the future. On the technical side, we will continue to explore additional signal-based behavior descriptors, e.g., lexical content, head pose, bodily gesture, etc., to better capture the multi-dimensional behavior feature space of the subjects and their interacting partners. Lastly, each of these various clinical instruments is often designed to measure a particular internal ability (e.g., cognitive, social, or communicative) of an ASD subject, computationally understand the relationship between them, e.g., how do measures of executive function as reflected in the CANTAB assessment related to the manifested behaviors in social contexts, would be an important research topic to better substantiate theory about autism such as theory of mind [64] and advance our understanding for a better development of clinical diagnostic instruments.

References

[1] J. Baio, L. Wiggins, D. L. Christensen, M. J. Maenner, J. Daniels, Z. Warren, M. Kurzius-Spencer, W. Zahorodny, C. R. Rosenberg, T. White, et al., Prevalence of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, united states, 2010, Morbidity and Mortality Weekly Report: Surveillance Summaries 63 (2) (2014) 1–21.

[2] C. Lord, E. H. Cook, B. L. Leventhal, D. G. Amaral, Autism spectrum disorders, Neuron 28 (2) (2000) 355–363.

[3] K. Rice, J. M. Moriuchi, W. Jones, A. Klin, Parsing heterogeneity in autism spectrum disorders: visual scanning of dynamic social scenes in school-aged children, Journal of the American Academy of Child & Adolescent Psychiatry 51 (3) (2012) 238–248.

[4] T. R. Insel, The nimh research domain criteria (rdoc) project: precision medicine for psychiatry, American Journal of Psychiatry 171 (4) (2014) 395–397.

[5] L. Wing, J. Gould, Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification, Journal of autism and developmental disorders 9 (1) (1979) 11–29.

[6] R. Landa, Early communication development and intervention for children with autism, Developmental Disabilities Research Reviews 13 (1) (2007) 16–25.

[7] G. W. H. Organization., The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines, Vol. 1, World Health Organization, 1992.

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

[9] C. Gillberg, C. Gillberg, M. Råstam, E. Wentz, The asperger syndrome (and high-functioning autism) diagnostic interview (asdi): a preliminary study of a new structured clinical interview, Autism 5 (1) (2001) 57–66.

[10] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, M. Rutter, The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism, Journal of autism and developmental disorders 30 (3) (2000) 205–223.

[11] S. Ozonoff, I. Cook, H. Coon, G. Dawson, R. M. Joseph, A. Klin, W. M. McMahon, N. Minshew, J. A. Munson, B. F. Pennington, et al., Performance on cambridge neuropsychological test automated battery subtests sensitive to frontal lobe function in people with autistic disorder: evidence from the collaborative programs of excellence in autism network, Journal of autism and developmental disorders 34 (2) (2004) 139–150.

[12] L. Bennetto, B. F. Pennington, S. J. Rogers, Intact and impaired memory functions in autism, Child development 67 (4) (1996) 1816–1835.

[13] S. Ozonoff, J. Jensen, Brief report: Specific executive function profiles in three neurodevelopmental disorders, Journal of autism and developmental disorders 29 (2) (1999) 171–177.

[14] M. Prior, W. Hoffmann, Brief report: Neuropsychological testing of autistic children through an exploration with frontal lobe tests, Journal of autism and developmental disorders 20 (4) (1990) 581–590.

[15] J. E. Russell, Autism as an executive disorder., Oxford University Press, 1997.

[16] E. L. Hill, Executive dysfunction in autism, Trends in cognitive sciences 8 (1) (2004) 26–32.

[17] T. W. Robbins, M. James, A. M. Owen, B. J. Sahakian, L. McInnes, P. Rabbitt, Cambridge neuropsychological test automated battery (cantab): a factor analytic study of a large sample of normal elderly volunteers, Dementia and Geriatric Cognitive Disorders 5 (5) (1994) 266–281.

[18] S. Ozonoff, M. South, J. N. Miller, Dsm-iv-defined asperger syndrome: Cognitive, behavioral and early history differentiation from high-functioning autism, Autism 4 (1) (2000) 29–46.

[19] S. D. Steele, N. J. Minshew, B. Luna, J. A. Sweeney, Spatial working memory deficits in autism, Journal of autism and developmental disorders 37 (4) (2007) 605–612.

[20] A. P. Association, Diagnostic and statistical manual of mental disorders (4th ed., text rev.), Washington, DC: Author, 1994.

[21] L. Wing, Asperger's syndrome: a clinical account, Psychological medicine 11 (1) (1981) 115–129.

[22] M. Ghaziuddin, L. Gerstein, Pedantic speaking style differentiates asperger syndrome from high-functioning autism, Journal of autism and developmental disorders 26 (6) (1996) 585–595.

[23] M. Ghaziuddin, Brief report: Should the dsm v drop asperger syndrome?, Journal of autism and developmental disorders 40 (9) (2010) 1146–1148.

[24] S. Narayanan, P. G. Georgiou, Behavioral signal processing: Deriving human behavioral informatics from speech and language, Proceedings of the IEEE 101 (5) (2013) 1203–1233.

[25] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, S. Narayanan, Signal processing and machine learning for mental health research and clinical applications [perspectives], IEEE Signal Processing Magazine 34 (5) (2017) 196–195.

[26] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, S. S. Narayanan, Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions, Computer Speech & Language 28 (2) (2014) 518–539.

[27] M. Reblin, R. E. Heyman, L. Ellington, B. R. Baucom, P. G. Georgiou, S. T. Vadaparampil, Everyday couples communication research: Overcoming methodological barriers with technology, Patient education and counseling 101 (3) (2018) 551–556.

[28] M. Nasir, B. Baucom, S. Narayanan, P. Georgiou, Towards an unsupervised entrainment distance in conversational speech using deep neural networks, arXiv preprint arXiv:1804.08782.

[29] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, S. Narayanan, Modeling therapist empathy and vocal entrainment in drug addiction counseling., in: INTERSPEECH, 2013, pp. 2861–2865.

[30] C.-P. Chen, X.-H. Tseng, S. S.-F. Gau, C.-C. Lee, Computing multimodal dyadic behaviors during spontaneous diagnosis interviews toward automatic categorization of autism spectrum disorder, in: Proc. Interspeech 2017, 2017, pp. 2361–2365. doi:10.21437/Interspeech.2017-563.
URL http://dx.doi.org/10.21437/Interspeech.2017-563

[31] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, S. Narayanan, Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist, in: Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[32] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, 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 57 (4) (2014) 1162–1177.

[33] W. Liu, M. Li, L. Yi, Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework, Autism Research 9 (8) (2016) 888–898.

[34] T. Zhou, W. Cai, X. Chen, X. Zou, S. Zhang, M. Li, Speaker diarization system for autism children's real-life audio data, in: Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on, IEEE, 2016, pp. 1–5.

[35] C. Leclère, M. Avril, S. Viaux-Savelon, N. Bodeau, C. Achard, S. Missonnier, M. Keren, R. Feldman,

M. Chetouani, D. Cohen, Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3d reconstruction, Translational psychiatry 6 (5) (2016) e816.

[36] B. Schuller, E. Marchi, S. Baron-Cohen, H. OReilly, P. Robinson, I. Davies, O. Golan, S. Friedenson, S. Tal, S. Newman, et al., Asc-inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions, in: Proceedings 1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2013) held in conjunction with the 8th Foundations of Digital Games 2013 (FDG)(B. Schuller, L. Paletta, and N. Sabouret, eds.), Chania, Greece, 2013.

[37] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, M. Plaza, Automatic intonation recognition for the prosodic assessment of language-impaired children, IEEE Transactions on Audio, Speech, and Language Processing 19 (5) (2011) 1328–1342.

[38] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013.

[39] H.-y. Lee, T.-y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, T.-L. Pao, Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition., in: INTERSPEECH, 2013, pp. 215–219.

[40] S. S.-F. Gau, C.-Y. Shang, Executive functions as endophenotypes in adhd: evidence from the cambridge neuropsychological test battery (cantab), Journal of Child Psychology and Psychiatry 51 (7) (2010) 838–849.

[41] Y.-L. Chien, S.-F. Gau, C.-Y. Shang, Y.-N. Chiu, W.-C. Tsai, Y.-Y. Wu, Visual memory and sustained attention impairment in youths with autism spectrum disorders, Psychological medicine 45 (11) (2015) 2263–2273.

[42] C. Hughes, J. Russell, T. W. Robbins, Evidence for executive dysfunction in autism, Neuropsychologia 32 (4) (1994) 477–492.

[43] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision, Sydney, Australia, 2013.
URL `http://hal.inria.fr/hal-00873267`

[44] D. Roy, C. K. Mohan, K. S. R. Murty, Action recognition based on discriminative embedding of actions using siamese networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3473–3477.

[45] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4305–4314.

[46] P. . Boersma, Praat, a system for doing phonetics by computer, Glot international 5:9/10 5.

[47] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE transactions on audio, speech, and language processing 17 (4) (2009) 582–596.

[48] J. Hillenbrand, R. A. Houde, Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech, Journal of Speech, Language, and Hearing Research 39 (2) (1996) 311–321.

[49] B. Halberstam, Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels, ORL 66 (2) (2004) 70–73.

[50] A. Mcallister, J. Sundberg, S. R. Hibi, Acoustic measurements and perceptual evaluation of hoarseness in children's voices, Logopedics Phoniatrics Vocology 23 (1) (1998) 27–38.

[51] M. Wilson, T. P. Wilson, An oscillator model of the timing of turn-taking, Psychonomic bulletin & review 12 (6) (2005) 957–968.

[52] Z. Warren, M. L. McPheeters, N. Sathe, J. H. Foss-Feig, A. Glasser, J. Veenstra-VanderWeele, A systematic review of early intensive intervention for autism spectrum disorders, Pediatrics 127 (5) (2011) e1303–e1311.

[53] L. A. LeBlanc, A. M. Coates, S. Daneshvar, M. H. Charlop-Christy, C. Morris, B. M. Lancaster, Using video modeling and reinforcement to teach perspective-taking skills to children with autism, Journal of applied behavior analysis 36 (2) (2003) 253–257.

[54] J. Brok, E. Barakova, Engaging autistic children in imitation and turn-taking games with multiagent system of interactive lighting blocks, Entertainment Computing-ICEC 2010 (2010) 115–126.

[55] M. Goudbeek, K. Scherer, Beyond arousal: Valence and potency/control cues in the vocal expression of emotion, The Journal of the Acoustical Society of America 128 (3) (2010) 1322–1336.

[56] C. Adams, J. Green, A. Gilchrist, A. Cox, Conversational behaviour of children with asperger syndrome and conduct disorder, Journal of Child Psychology and Psychiatry 43 (5) (2002) 679–690.

[57] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE Transactions on Affective Computing 7 (2) (2016) 190–202.

[58] N. Matsuura, M. Ishitobi, S. Arai, K. Kawamura, M. Asano, K. Inohara, T. Narimoto, Y. Wada, M. Hiratani, H. Kosaka, Distinguishing between autism spectrum disorder and attention deficit hyperactivity disorder by using behavioral checklists, cognitive assessments, and neuropsychological test battery, Asian journal of psychiatry 12 (2014) 50–57.

[59] A. Kushki, J. Brian, A. Dupuis, E. Anagnostou, Functional autonomic nervous system profile in children with autism spectrum disorder, Molecular autism 5 (1) (2014) 39.

[60] S. Ozonoff, D. L. Strayer, Further evidence of intact working memory in autism, Journal of autism and developmental disorders 31 (3) (2001) 257–263.

[61] H. B. Garretson, D. Fein, L. Waterhouse, Sustained attention in children with autism, Journal of autism and developmental disorders 20 (1) (1990) 101–114.

[62] A. B. Sereno, S. C. Amador, Attention and memory-related responses of neurons in the lateral intraparietal area during spatial and shape-delayed match-to-sample tasks, Journal of neurophysiology 95 (2) (2006) 1078–1098.

[63] S.-F. Chen, Y.-L. Chien, C.-T. Wu, C.-Y. Shang, Y.-Y. Wu, S. Gau, Deficits in executive functions among youths with autism spectrum disorders: an age-stratified analysis, Psychological medicine 46 (8) (2016) 1625–1638.

[64] S. Baron-Cohen, Mindblindness: An essay on autism and theory of mind, MIT press, 1997.

Table 2: Coding in ADOS, four catagories, subsets of the four catagories and their abbreviation

Catagory	abbreviation	description
Language and Communication	UOTH	Use of Other's Body to Communicate
	VOC	Frequency of Vocalization Directed to Others
	ASOV	Amount of Social OverturesjMaintenance of Attention
	PNT	Pointing
	STER	Stereotyped ldiosyncratic Use of Words or Phrases
	DGES	Descriptive conventional Instrumental or Informational Gestures
	GES	Gestures
	IECHO	Immediate Echolalia
	SPAB	Speech Abnormalities Associated With Autism
	CONV	Conversation
	REPT	Reporting of Events
	OINF	Offers Information
	EGES	Emphatic or Emotional Gestures
Reciprocal Social lnteraction	SHO	Showing
	IJA	Spontaneous Initiation of Joint Attention
	RJA	Response to Joint Attention
	EXP	Facial Expressions Directed to Others
	QSOV	Quality of Social Overtures
	EYE	UnusualEye Contact
	ENJ	Shared Enjoyment in Interaction
	GAZE	Integration of Gaze and Other Behaviors During Social Overtures
	ARSC	amount of reciprocal social communication
	OQR	Overall Quality of Rapport
	QSR	Quality of SocialResponse
	INS	Insight
	EMP	Empathy/Comments on Others' Emotions
	RESP	Responsibility
Play + Imagination/Creativity	PLAY	Functional Play With Objects
	IMAG	ImaginationjCreativity
Stereotyped Behavior and Restricted Interests	MAN	Hand and Finger and Other Complex Mannerisms
	RINT	Unusually Repetitive Interests or Stereotyped Behaviors
	RITL	Compulsions or Rituals
	SINT	Unusual Sensory Interest in Play Material Person
	XINT	Excessive Interest in Unusual or Highly Specific Topics or Objects

Table 5: Classification UAR using single modality behavior features. Abbreviated symbols' explanation F: functional, μ: mean, σ: standard deviation, The abbreviation of the task: AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism

Feature_name	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$^{investquest}_{invest}$]	0.51/0.54/0.57	0.57/0.46/0.58	0.61/0.61/0.56	0.4/0.4/0.36
μ-[NBAE$^{gap}_{invest}$]	0.59/0.61/0.67	0.64/0.61/0.68	0.7/0.69/0.65	0.51/0.52/0.43
μ-[NBAE$^{partresp}_{invest}$]	0.54/0.54/0.55	0.46/0.46/0.61	0.5/0.46/0.56	0.33/0.3/0.38
μ-[NBAE$^{turntaking}_{invest}$]	0.55/0.54/0.56	0.63/0.68/0.76	0.64/0.62/0.62	0.38/0.37/0.42
μ-[NBAE$^{investquest}_{part}$]	0.41/0.42/0.6	**0.71**/0.63/0.72	0.67/0.67/0.71	0.39/0.39/0.45
μ-[NBAE$^{gap}_{part}$]	0.63/0.54/0.65	0.4/0.42/0.61	0.61/0.59/0.62	0.39/0.39/0.39
μ-[NBAE$^{partresp}_{part}$]	0.42/0.43/0.54	0.54/0.54/0.54	0.57/0.52/0.67	0.32/0.32/0.4
μ-[NBAE$^{turntaking}_{part}$]	0.54/0.56/0.59	0.54/0.54/0.55	0.48/0.45/0.53	0.38/0.35/0.37
μ-[NBAE$^{investquest}_{inter}$]	0.51/0.6/0.72	0.59/0.64/0.69	0.54/0.49/0.6	0.34/0.33/0.43
μ-[NBAE$^{partresp}_{inter}$]	0.51/0.43/0.52	0.63/0.55/0.66	0.54/0.49/0.57	0.34/0.35/0.37
μ-[NBAE$^{gap}_{inter}$]	0.62/0.66/0.64	0.65/0.64/0.73	0.51/0.46/0.62	0.34/0.35/0.45
μ-[NBAE$^{turntaking}_{inter}$]	**0.65**/0.65/0.59	0.59/0.63/0.61	0.57/0.59/0.61	0.34/0.35/0.39
μ-[Pitch$^{investquest}_{invest}$]	0.54/0.53/0.56	0.46/0.44/0.51	0.4/0.43/0.53	0.34/0.33/0.34
μ-[Pitch$^{partresp}_{part}$]	0.42/0.42/0.57	0.55/0.55/0.71	0.48/0.54/0.54	0.27/0.29/0.43
μ-[Pitch$^{investquest/partresp}_{inter}$]	0.5/0.56/0.56	0.47/0.53/0.66	0.5/0.56/0.61	0.33/0.39/0.43
σ-[Pitch$^{investquest}_{invest}$]	0.45/0.42/0.64	0.54/0.57/0.7	0.45/0.53/0.57	0.26/0.26/0.39
σ-[Pitch$^{partresp}_{part}$]	0.49/0.56/0.67	0.63/0.61/0.63	**0.75**/0.71/0.71	0.39/0.46/0.44
σ-[Pitch$^{investquest/partresp}_{inter}$]	0.53/0.5/0.57	0.48/0.46/0.53	0.45/0.51/0.63	0.35/0.33/0.37
μ-[Intensity$^{investquest}_{invest}$]	0.55/0.58/0.53	0.42/0.45/0.67	0.45/0.43/0.57	0.34/0.34/0.37
μ-[Intensity$^{partresp}_{part}$]	0.46/0.51/0.59	0.51/0.47/0.57	0.55/0.47/0.55	0.31/0.27/0.42
μ-[Intensity$^{investquest/partresp}_{inter}$]	0.45/0.51/0.55	0.45/0.56/0.47	0.44/0.48/0.58	0.26/0.35/0.36
σ-[Intensity$^{investquest}_{invest}$]	0.57/0.67/0.6	0.62/0.62/0.55	0.66/0.67/0.64	0.4/0.46/0.4
σ-[Intensity$^{partresp}_{part}$]	0.55/0.54/0.51	0.47/0.41/0.51	0.58/0.68/0.58	0.36/0.33/0.36
σ-[Intensity$^{investquest/partresp}_{inter}$]	0.59/0.64/0.58	0.64/0.68/0.71	0.53/0.57/0.62	0.41/0.47/0.43
μ-[HNR$^{investquest}_{invest}$]	0.57/0.58/0.7	0.48/0.57/0.7	0.49/0.61/0.63	0.34/0.38/0.39
μ-[HNR$^{partresp}_{part}$]	0.49/0.62/0.62	0.48/0.55/0.48	0.61/0.58/0.59	0.34/0.37/0.42
σ-[HNR$^{investquest}_{invest}$]	0.39/0.56/0.54	0.5/0.51/0.59	0.67/0.57/0.56	0.3/0.38/0.33
σ-[HNR$^{partresp}_{part}$]	0.5/0.53/0.65	0.55/0.55/0.69	0.51/0.57/0.65	0.38/0.38/0.48
μ-[Jitter$^{investquest}_{invest}$]	0.43/0.54/0.55	0.5/0.5/0.53	0.6/0.53/0.54	0.23/0.33/0.35
μ-[Jitter$^{partresp}_{part}$]	0.39/0.52/0.49	0.48/0.56/0.61	0.38/0.49/0.53	0.24/0.32/0.37
μ-[Shimmer$^{investquest}_{invest}$]	0.49/0.51/0.62	0.61/0.63/0.72	0.64/0.57/0.57	0.33/0.33/0.34
μ-[Shimmer$^{partresp}_{part}$]	0.29/0.18/0.49	0.57/0.57/0.5	0.56/0.56/0.56	0.33/0.33/0.33
Duration$^{investquest}_{intra}$	0.5/0.52/0.52	0.47/0.56/0.51	0.49/0.55/0.59	0.2/0.32/0.38
Duration$^{gap}_{intra}$	0.63/0.6/0.63	0.56/0.73/0.75	0.61/0.65/0.75	0.44/0.5/0.49
Duration$^{partresp}_{intra}$	0.49/0.49/0.54	0.45/0.48/0.57	0.44/0.5/0.51	0.29/0.29/0.36
Duration$^{turntaking}_{intra}$	0.49/0.55/0.68	0.49/0.57/0.64	0.61/0.61/0.71	0.44/0.44/0.43
Duration$^{gap/investquest}_{inter}$	0.56/0.57/0.57	0.5/0.49/0.49	0.58/0.53/0.5	0.37/0.36/0.35
Duration$^{partresp/gap}_{inter}$	0.51/0.56/0.65	0.53/0.53/0.67	0.63/0.61/0.59	0.35/0.38/0.43
Duration$^{gap/investquest}_{inter}$	0.61/0.55/0.58	0.54/0.51/0.68	0.7/0.57/0.63	**0.53**/0.39/0.5
Duration$^{investquest/gap}_{inter}$	0.49/0.54/0.75	0.54/0.62/0.62	0.6/0.64/0.68	0.23/0.4/0.61
Duration$^{partresp/investquest}_{inter}$	0.3/0.61/0.65	0.43/0.57/0.7	0.52/0.71/0.65	0.2/0.53/0.46

Table 6: Multimodal classification with classifier SVM (left) and randomforest (right) on the designed tasks. Bolded value means its value is higher than baseline (ADOS Communication, Social Reciprocity), and the highest value in each task is highlighted in red color. The meanings of abbreviations are listed below, AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism

F-(action,acoustic, turn-taking)	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$_i^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.54/0.62	0.66/0.75	0.66/0.63	0.44/0.44
μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$	0.41/0.54	0.43/0.75	0.51/0.63	0.23/0.49
μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.47/0.56	0.69/0.63	0.65/0.67	0.46/0.45
μ-[NBAE$_i^{pr}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.57/0.64	0.55/0.65	0.62/0.59	0.4/0.43
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.7/0.65	0.58/0.68	0.55/0.64	0.54/0.44
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{intra}^{tt}$	0.51/0.64	0.71/0.68	0.63/0.68	0.46/0.43
μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$	0.44/0.62	0.57/0.66	0.71/0.66	0.49/0.46
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{iq}$	0.48/0.56	0.62/0.64	0.67/0.66	0.35/0.41
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{gap}$	0.45/0.64	0.78/0.67	0.7/0.71	0.41/0.46
μ-[NBAE$_{inter}^{gap}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{pr}$	0.67/0.62	0.63/0.69	0.45/0.54	0.35/0.35
μ-[NBAE$_{inter}^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.59/0.62	0.59/0.65	0.55/0.59	0.39/0.41

Table 7: Left: Classification UAR using multimodal behavior on the 52 available subjects. Right: Classification UAR after fusing behavior features with CANTAB measures. The meanings of abbreviations are listed below μ: mean, σ: standard deviation, AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism. The classifier is chosen to be SVC

F-(action,acoustic, turn-taking)	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$_i^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.47/0.59	0.69/0.65	0.65/0.64	0.46/0.4
μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$	0.57/0.64	0.55/0.71	0.62/0.69	0.4/0.41
μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.7/0.6	0.58/0.63	0.55/0.67	0.54/0.44
μ-[NBAE$_i^{pr}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.51/0.57	0.71/0.68	0.63/0.63	0.46/0.4
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.44/0.6	0.57/0.65	0.71/0.64	0.49/0.42
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{intra}^{tt}$	0.48/0.57	0.62/0.64	0.67/0.64	0.35/0.41
μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$	0.45/0.6	0.78/0.64	0.7/0.66	0.41/0.41
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{iq}$	0.67/0.58	0.63/0.63	0.45/0.62	0.35/0.43
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{gap}$	0.59/0.6	0.59/0.64	0.55/0.65	0.39/0.4
μ-[NBAE$_{inter}^{gap}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{pr}$	0.51/0.59	0.6/0.66	0.55/0.65	0.4/0.41
μ-[NBAE$_{inter}^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.46/0.58	0.69/0.64	0.59/0.62	0.42/0.43
Stepwise(CANTAB)	0.76	0.81	0.74	0.54

Table 8: Left: Classification UAR using multimodal behavior on the 52 available subjects. Right: Classification UAR after fusing behavior features with CANTAB measures. The meanings of abbreviations are listed below μ: mean, σ: standard deviation, AD: autism, AS: Asperger's syndrome, HFA: High-functioning autism. The classifier is chosen to be random forest

F-(action,acoustic, turn-taking)	AD vs AS	AS vs HFA	AD vs HFA	AD vs AS vs HFA
μ-[NBAE$_i^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.56/0.71	0.63/0.62	0.67/0.76	0.45/0.47
μ-[NBAE$_i^{pr}$] +μ-[HNR$_p^{pr}$] +Duration$_{intra}^{gap}$	0.64/0.75	0.65/0.62	0.59/0.76	0.43/0.43
μ-[NBAE$_i^{pr}$] +μ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.65/0.76	0.68/0.64	0.64/0.75	0.44/0.44
μ-[NBAE$_i^{pr}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.64/0.74	0.68/0.63	0.68/0.75	0.43/0.44
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{inter}^{gap/iq}$	0.62/0.77	0.66/0.68	0.66/0.76	0.46/0.46
μ-[NBAE$_p^{iq}$] +σ-[Intensity$_i^{iq}$] +Duration$_{intra}^{tt}$	0.56/0.78	0.64/0.64	0.66/0.78	0.41/0.47
μ-[NBAE$_p^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{iq}$	0.64/0.72	0.67/0.64	0.71/0.77	0.46/0.46
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{iq}$	0.62/0.71	0.69/0.65	0.54/0.75	0.35/0.39
μ-[NBAE$_{inter}^{gap}$] +μ-[Intensity$_i^{iq}$] +Duration$_{intra}^{gap}$	0.62/0.72	0.65/0.66	0.59/0.75	0.41/0.42
μ-[NBAE$_{inter}^{gap}$] +σ-[Pitch$_p^{pr}$] +Duration$_{intra}^{pr}$	0.6/0.73	0.67/0.65	0.69/0.76	0.43/0.45
μ-[NBAE$_{inter}^{iq}$] +σ-[Pitch$_p^{pr}$] +Duration$_{inter}^{gap/iq}$	0.62/0.77	0.7/0.71	0.66/0.76	0.46/0.44
Stepwise(CANTAB)	0.76	0.81	0.74	0.54