Investigating the Interplay Between Affective, Phonatory and Motoric Subsystems in Autism Spectrum Disorder Using a Multimodal Dialogue Agent

Hardik Kothare\textsuperscript{1,2}, Vikram Ramanarayanan\textsuperscript{1,2}, Oliver Roesler\textsuperscript{1}, Michael Neumann\textsuperscript{1}, Jackson Liscombe\textsuperscript{1}, William Burke\textsuperscript{1}, Andrew Cornish\textsuperscript{1}, Doug Habberstad\textsuperscript{3}, Alaa Sakallah\textsuperscript{2,3}, Sara Markuson\textsuperscript{2,3}, Seemran Kansara\textsuperscript{2,3}, Afik Faerman\textsuperscript{2,3}, Yasmine Bensidi-Slimane\textsuperscript{2,3}, Laura Fry\textsuperscript{2,3}, Saige Portera\textsuperscript{2,3}, David Suendermann-Oeft\textsuperscript{1}, David Pautler\textsuperscript{1}, Carly Demopoulos\textsuperscript{2}

\textsuperscript{1} Modality.AI, Inc.
\textsuperscript{2} University of California, San Francisco
\textsuperscript{3} Palo Alto University

\textsuperscript{4} University of California, Berkeley

\texttt{<hardik.kothare,vikram.ramanarayanan>@modality.ai}

Abstract

We explore the utility of an on-demand multimodal conversational platform in extracting speech and facial metrics in children with Autism Spectrum Disorder (ASD). We investigate the extent to which these metrics correlate with objective clinical measures, particularly as they pertain to the interplay between the affective, phonatory and motoric subsystems. 22 participants diagnosed with ASD engaged with a virtual agent in conversational affect production tasks designed to elicit facial and vocal affect. We found significant correlations between vocal pitch and loudness extracted by our platform during these tasks and accuracy in recognition of facial and vocal affect, assessed via the Diagnostic Analysis of Nonverbal Accuracy-2 (DANVA-2) neuropsychological task. We also found significant correlations between jaw kinematic metrics extracted using our platform and motor speed of the dominant hand assessed via a standardised neuropsychological finger tapping task. These findings offer preliminary evidence for the usefulness of these audiovisual analytic metrics and could help us better model the interplay between different physiological subsystems in individuals with ASD.

Index Terms: multimodal conversational agent, autism spectrum disorder, affect production, motor coordination, audiovisual analytics.

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that manifests as deficits in social communication and interaction [1]. Epidemiological surveys estimate the prevalence of ASD in the United States at 18.5 per 1,000 children [2]. Autism diagnosis rates have increased due to an increase in awareness and access to resources. In the state of California, the diagnosed autism incidence rate increased by 612% in 2003 [3]. Moreover, the coordination of facial and vocal expressions during emotional speech production is less coordinated in ASD [20]. Due to such cross-domain atypicalities in ASD, a multimodal framework approach has been suggested to provide useful quantitative insights in early diagnosis and categorisation of ASD [21, 22, 23]. In this paper, we explore the interplay between multiple neuropsychological subsystems through examination of automatically-extracted voice and facial metrics during a novel conversational task and a comprehensive set of clinically-validated objective measures. We probe the feasibility of a scalable, low-cost and remotely-administrable multimodal conversational platform for this purpose. Such technology could potentially assist clinicians and researchers in gathering relevant diagnostic information and in longitudinally monitoring children with developmental disorders.

2. Multimodal Conversational Platform

NEMSI or NEurological and Mental health Screening Instrument is a cloud-based multimodal dialogue technology in which participants engage in a conversation with a virtual agent and report on their social, emotional and communication abilities. Data analytics modules automatically extract relevant speech and facial features from the captured audiovisual data in real time and store them in a database. These metrics, with a detailed participant-wise, session-wise and task-wise breakdown, can be accessed by researchers via a user-friendly dashboard. A more detailed description of NEMSI, conversational protocols and a schematic of the software and hardware architecture of the system can be found in our previous publications [24, 25, 26].
3. Methods

The current study was conducted on site at the University of California, San Francisco and was approved by the university’s Institutional Review Board. Informed consent was obtained from the participants’ guardians along with a written assent from the participants. Data from 22 participants (10 females, mean age: 11.37 ± 2.47 years, see Figure 1) diagnosed with ASD by a licensed clinical psychologist was included in the analysis.

3.1. Standardised Clinical Instruments

All participants / guardians underwent a thorough neuropsychological evaluation battery including the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [27], the Autism Diagnostic Interview - Revised [28], the Behavior Assessment System for Children, Third Edition (BASC-3) [29], the Diagnostic Analysis of Nonverbal Accuracy (DANVA) [30, 31] and NEPSY-II [32]. The ADOS-2 scores for 11 participants were collected during the COVID-19 pandemic and were deemed invalid because the participants wore a mask during the assessment. These scores were excluded from the analysis. Additionally, participants also took part in a motor skills and fine motor dexterity assessment, including bilateral manual grip strength assessed with a Lafayette Hand Dynamometer Model J00105 [33], finger tapping speed assessed with a standard board-mounted finger tapper with counter [34], and the Grooved Pegboard Test [35]. All scores were age scaled based on the referenced norms.

3.2. Affect Production Task

The Affect Production Task (APT) was administered via Modality AI’s NEMSI platform. Participants were prompted to produce four emotions (happy, sad, angry and afraid) through five tasks:

1. Monosyllabic Emotion Production (8 turns, 2 turns per emotion): a video stimulus prompted the users to say “oh” in a way that conveyed the specified emotion during every turn;
2. Sentence-length Emotion Production (8 turns, 2 turns per emotion): a video stimulus prompted the users to say “I’ll be right back” in a way that conveyed the specified emotion during every turn;
3. Emotion-eliciting Situations (16 turns, 4 turns per emotion): the virtual agent narrated a situation and a corresponding emotional response at the end of which a picture depicting the situation was shown and the user was prompted to say “oh” in a way that conveyed the specified emotion that was situationally appropriate;
4. Monosyllabic Repetition (16 turns, 4 turn per emotion): an audio recording of “oh” in one of the four emotions was played during every turn and the user was asked to repeat the monosyllabic production in the same emotional inflection;
5. Facial Repetition (8 turns, 2 turns per emotion): a video prompted the user to make a face representing one of the four emotions like the person in the video during every turn.

3.3. Analytics

Speech metrics (Fundamental Frequency in Hz, Articulation Loudness in dB) and facial metrics (Eyebrow Height in pixels, Eye Openness in pixels, Lip Aperture in pixels, Mouth Surface Area in pixels, Jaw Velocity in pixels/frames and Jaw Acceleration pixels/frames) were calculated in real time [25] and displayed on a dashboard accessible to the clinicians and the researchers involved in the study. Facial metrics in pixels were normalised within every subject by dividing the values by the inter-lachrymal distance in pixels for each subject.

3.4. Statistical Methods

Linear correlations were run between measures from standardised neuropsychological assessments and the speech and facial metrics derived from our platform. To control for false positives due to multiple comparison, the Benjamini-Hochberg procedure [36] was used to adjust the p-values using a false discovery rate of $\alpha = 0.05$. Findings that survived this adjusted significance threshold were included in this paper.

4. Observations

4.1. Language Level and IQ

The Clinical Evaluation of Language Fundamentals- Fifth Edition (CELF-5) [37] and the Weschler Intelligence Scale for Children - Full Scale IQ (WISC - FSIQ) [38] scores for the cohort are reported in Table 1.

4.2. Speech Metrics

We found significant positive correlations between the DANVA Adult Receptive Paralanguage Subscore (APSS) and the fundamental frequency of the participants’ speech during the Emotion-Eliciting Situation subtask in the APT (Happy: $r^2 = 0.41$ and Sad: $r^2 = 0.54$) and also in the Monosyllabic Emotion Production subtask. (Happy: $r^2 = 0.56$ and Angry: $r^2 = 0.44$). We also found positive correlations between the DANVA Child Receptive Paralanguage Subscore (CPSS) and the fundamental frequency of the participants’ speech during the Emotion-
### Table 1: Language level and IQ measures in the cohort

<table>
<thead>
<tr>
<th>Measure</th>
<th>Range</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELF-5 Core Language Score</td>
<td>59 - 143</td>
<td>100.77 ± 22.06</td>
</tr>
<tr>
<td>CELF-5 Expressive Language Index</td>
<td>59 - 131</td>
<td>100.09 ± 20.48</td>
</tr>
<tr>
<td>CELF-5 Receptive Language Index</td>
<td>58 - 141</td>
<td>100.73 ± 23.08</td>
</tr>
<tr>
<td>WISC - FSIQ</td>
<td>50 - 136</td>
<td>102.95 ± 19.79</td>
</tr>
</tbody>
</table>

### Table 2: Speech acoustic measures of fundamental frequency (F0) and articulation loudness extracted automatically correlate with the participants’ accuracy at recognition of facial and vocal affect in audio recordings of adults and children

<table>
<thead>
<tr>
<th>Objective Clinical Measure</th>
<th>Extracted Analytics</th>
<th>Prompt</th>
<th>Sample Size</th>
<th>$r^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DANVA APSS</td>
<td>Fundamental Frequency</td>
<td>Emotion-Eliciting Situation HAPPY</td>
<td>15</td>
<td>0.41</td>
<td>0.0099</td>
</tr>
<tr>
<td>DANVA APSS</td>
<td>Fundamental Frequency</td>
<td>Monosyllabic Emotion Production HAPPY</td>
<td>19</td>
<td>0.56</td>
<td>0.0002</td>
</tr>
<tr>
<td>DANVA APSS</td>
<td>Fundamental Frequency</td>
<td>Emotion-Eliciting Situation SAD</td>
<td>17</td>
<td>0.54</td>
<td>0.0007</td>
</tr>
<tr>
<td>DANVA APSS</td>
<td>Fundamental Frequency</td>
<td>Monosyllabic Emotion Production ANGRY</td>
<td>18</td>
<td>0.44</td>
<td>0.0002</td>
</tr>
<tr>
<td>DANVA CPSS</td>
<td>Fundamental Frequency</td>
<td>Emotion-Eliciting Situation HAPPY</td>
<td>16</td>
<td>0.55</td>
<td>0.0016</td>
</tr>
<tr>
<td>DANVA CPSS</td>
<td>Fundamental Frequency</td>
<td>Monosyllabic Emotion Production HAPPY</td>
<td>19</td>
<td>0.48</td>
<td>0.0010</td>
</tr>
<tr>
<td>DANVA CPSS</td>
<td>Fundamental Frequency</td>
<td>Emotion-Eliciting Situation SAD</td>
<td>17</td>
<td>0.45</td>
<td>0.0034</td>
</tr>
<tr>
<td>DANVA CPSS</td>
<td>Fundamental Frequency</td>
<td>Monosyllabic Repetition AFRAID</td>
<td>16</td>
<td>0.36</td>
<td>0.0108</td>
</tr>
<tr>
<td>DANVA CPSS</td>
<td>Articulation Loudness</td>
<td>Sentence-length Emotion Production HAPPY</td>
<td>16</td>
<td>0.34</td>
<td>0.0180</td>
</tr>
</tbody>
</table>

### Table 3: Jaw kinematic measures (velocity and acceleration) extracted automatically correlate with motor speed of the dominant hand

<table>
<thead>
<tr>
<th>Objective Clinical Measure</th>
<th>Extracted Analytics</th>
<th>Prompt</th>
<th>Sample Size</th>
<th>$r^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger Tapping Dominant Hand</td>
<td>Maximum Jaw Velocity</td>
<td>Monosyllabic Repetition SAD</td>
<td>13</td>
<td>0.44</td>
<td>0.0139</td>
</tr>
<tr>
<td>Finger Tapping Dominant Hand</td>
<td>Maximum Jaw Velocity</td>
<td>Monosyllabic Repetition AFRAID</td>
<td>13</td>
<td>0.35</td>
<td>0.0317</td>
</tr>
<tr>
<td>Finger Tapping Dominant Hand</td>
<td>Maximum Jaw Velocity</td>
<td>Monosyllabic Repetition SAD</td>
<td>13</td>
<td>0.35</td>
<td>0.0345</td>
</tr>
<tr>
<td>Finger Tapping Dominant Hand</td>
<td>Maximum Jaw Acceleration</td>
<td>Emotion-Eliciting Situation AFRAID</td>
<td>13</td>
<td>0.42</td>
<td>0.0164</td>
</tr>
</tbody>
</table>

Eliciting Situation subtask (Happy: $r^2 = 0.55$ and Sad: $r^2 = 0.45$), the Monosyllabic Emotion Production subtask (Happy: $r^2 = 0.48$) and the Monosyllabic Repetition subtask (Angry: $r^2 = 0.36$). Fundamental frequency of speech during all these tasks did not correlate with participants’ age so the effect of age on these correlations can be ruled out. Additionally, the DANVA CPSS score was also positively correlated with Articulation Loudness during the Sentence-length Emotion Production subtask (Happy: $r^2 = 0.34$). Please refer to Table 2 for more details.

#### 4.3. Facial Metrics

Motor speed of the dominant hand assessed using a finger tapping task was found to be positively correlated with participants’ maximum jaw velocity during the Monosyllabic Repetition subtask in the APT (Sad: $r^2 = 0.43$ and Afraid: $r^2 = 0.37$) and during the Emotion-Eliciting Situation subtask (Afraid: $r^2 = 0.36$). Motor speed of the dominant hand was also positively correlated with maximum jaw acceleration during the Emotion-Eliciting Situation subtask (Afraid: $r^2 = 0.42$). Please refer to Table 3 for more details.

### 5. Discussion

In this study, we found that speech metrics derived by NEMSI, a cloud-based multimodal dialogue technology, are significantly correlated with recognition of vocal affect assessed via the Diagnostic Analysis of Nonverbal Accuracy (DANVA) task. We also found correlations between jaw kinematics and motor speed of the dominant hand as assessed by a standardised finger tapping task.

The DANVA receptive paralanguage subtasks assess the participants’ ability to identify vocal emotional communication, with higher scores indicating competence in recognition of vocal affect. Participants who were more competent in identifying the emotion from adult and child voice recordings produced emotional speech (monosyllabic “oh”) with a higher fundamental frequency. A previous study has found an increased pitch range in high-functioning autism [9]. Participants with better DANVA CPSS scores also produced louder sentence-length happy emotional speech. Increased articulation loudness is indeed associated with happy emotional speech [39]. Our findings may suggest that high-performing participants on the vocal affect recognition task may possess greater ability to modulate the acoustic properties of emotional speech whereas more impaired participants are unable to do so. This may suggest that there is a relationship between the recognition and production of the acoustic properties that are relevant for signifying emotion. Indeed, previous research has identified a relationship between basic auditory processing and vocal affect recognition [40, 41, 42].

The correlations between jaw movement and motor dexterity of the dominant hand may be a manifestation of global motor impairments in individuals with ASD [43, 44]. Motor deficits in ASD have been attributed to abnormal sensorimotor integration [45, 46] due to increased sensory noise [47] or dysfunctional representation of internal models of action [48]. A previous study has found no differences in finger-tapping accuracy between children with ASD and typically-developing children but there was increased variability in temporal processing...
parameters in the ASD group [49]. Correlations between jaw movement and finger tapping suggest global altered temporal processing in ASD which may also explain poor cross-modal coordination during emotional speech production [20]. Indeed, prior work has suggested a coupling between speech motor coordination and fine motor skill systems in ASD [50].

Our preliminary results indicate that speech and facial metrics extracted from our APT task administered using Modality.AI’s NEMSI platform can tap into deficits of various physiological subsystems in children with ASD. The current study was conducted in a laboratory setting but the conversational platform can also be accessed remotely via a weblink [26]. Thus, our technology provides the opportunity for clinicians and researchers to monitor children with developmental disorders in a non-clinical, non-laboratory setting. Since deficits in emotion recognition have been observed cross-culturally in individuals with ASD [51], our technology can also be leveraged to compile a global corpus of speech and facial metrics during affect recognition and production in ASD. A conversational task where participants are prompted to produce a monosyllabic “oh” may also prove useful in a low-functioning population. Furthermore, it has been shown that children with ASD can be trained to produce qualitatively-improved emotional facial expressions [52]. Remote monitoring may also facilitate a longitudinal, quantitative analysis of such training paradigms.

Future studies would involve collecting data from larger cohorts of children with ASD along with data from typically-developing children to examine whether differences in vocal and facial affect production between the two groups can be captured through automatically-extracted metrics.

6. Acknowledgements

This study was supported by the National Institutes of Health grant K23 DC016637 and Autism Speaks grant 11637 awarded to Carly Demopoulos.

7. References


