Atypical speech acoustics and jaw kinematics during affect production in children with Autism Spectrum Disorder assessed by an interactive multimodal conversational platform

Hardik Kothare1, Vikram Ramanarayanan1,2, Oliver Roessler1, Michael Neumann1, Jackson Liscombe1, William Burke1, Andrew Cornish1, Doug Habberstad1, Brandon Kopal1, Alison Bai2, Yelena Markiv2, Levi Cole2, Sara Markuson1, Yasminet Bensidi-Slimane1, Alaak Sakkallah1, Katherine Brogan1, Linnea Lampinen1, Sara Skiba1, David Suendermann-Oeft1, David Pautler1, Carly Dermopoulos1

1Modality.AI, Inc., 2University of California, San Francisco

Introduction

Objective: Identify audiovisual speech markers that show significant differences between children with Autism Spectrum Disorder (ASD) and controls.

Task: A novel affect production task conducted by a virtual dialogue agent via a cloud-based multimodal conversational platform.

Implications: Objective audiovisual metrics of speech motor control during affect production in ASD may be used as diagnostic aids and in tracking the outcome of potential interventions.

Methods and Materials

- 44 participants with ASD (16 female, mean age = 11.74 ± 2.56 years) and 17 controls (8 female, mean age = 12.80 ± 2.59 years) completed an interactive session between December 2019 and February 2022 using a cloud-based multimodal dialogue platform (illustration in Figure 1).

- Participants were asked to produce one of four emotions: Happy, Sad, Angry, Afraid through the following tasks:
  - Task 1: Repeat the monosyllable "oh" after a video stimulus
  - Task 2: Repeat the monosyllable "oh" after an audio stimulus
  - Task 3: Produce the monosyllable "oh" after a situation narration followed by a picture stimulus
  - Task 4: Repeat the sentence "I’ll be right back" after a video stimulus

- Facial metrics were normalised for each participant by the inter-caruncular distance between the eyes. Automatically-extracted speech acoustic and facial kinematic metrics were further normalised by gender.

- Non-parametric Kruskal-Wallis tests were performed to investigate differences between ASD and controls.

Table 1. Automatically extracted acoustic & visual measures.

<table>
<thead>
<tr>
<th>Acoustic measures</th>
<th>Visual measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental Frequency (F0): Minimum value (Hz) and timepoint (s), Maximum value (Hz) and timepoint (s), Mean (Hz), Standard Deviation (Hz)</td>
<td>velocity, acceleration, and jerk of lower lip and jaw center, lip aperture, lip width, eye opening, vertical eyebrow displacement, eye blinks, area of the mouth, symmetry ratio of the mouth area</td>
</tr>
<tr>
<td>Cepstral Peak Prominence (CPP in dB)</td>
<td></td>
</tr>
<tr>
<td>Harmonics-to-Noise Ratio (HNR in dB)</td>
<td></td>
</tr>
<tr>
<td>Articulation rate (in s, excluding pauses) and speaking time (in s, including pauses)</td>
<td></td>
</tr>
<tr>
<td>Articulation rate and speaking rate (words per minute)</td>
<td></td>
</tr>
<tr>
<td>Percent pause duration (%)</td>
<td></td>
</tr>
<tr>
<td>Signal-to-noise ratio (SNR in dB)</td>
<td></td>
</tr>
<tr>
<td>Articulation intensity (dB)</td>
<td></td>
</tr>
<tr>
<td>Jitter and shimmer (%)</td>
<td></td>
</tr>
</tbody>
</table>

Results and Discussion

- A variety of metrics showed statistically significant differences between the ASD cohort and controls (Figure 2).

  - **Jaw kinematics:**
    - The ASD cohort exhibited greater velocity, acceleration and jerk of the jaw only for two of the four emotions - angry and afraid - and only in two of the four tasks, i.e. when participants were asked to repeat monosyllabic or sentential speech after a video stimulus. Greater variance of these jaw kinetic metrics in the ASD cohort, as evaluated by Fligner-Killeen tests.
    - This suggests exaggerated jaw movement while mimicking speech with negative emotions from a video stimulus but not when affect production is elicited via a picture stimulus or repetition of an audio stimulus.
    - **Spectral metrics:**
      - Larger formant frequency values of the monosyllabic vowel /a/ in ASD, elicited by a picture stimulus or the audio repetition of sad, afraid and angry emotions.
      - Larger maximum F0 in ASD during afraid sentential repetition.
    - All the above differences showed a statistically significant difference at an alpha threshold of 0.05 and were controlled for false discovery rate.

Conclusions

- The findings point towards exaggerated and variable speech motor control in ASD during repetition of emotional speech only when the production is cued via a video.

- **Acoustic properties of emotional speech in ASD are atypical.**

- These differences are specific to certain emotions providing a novel insight into the atypical production of vocal and facial affect during emotional speech in ASD.

Table 1. Automatically extracted acoustic & visual measures.

References