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

Hardik Kothare[1], Vikram Ramanarayanan[1,2], Oliver Roesler[1], Michael Neumann[1], Jackson Liscombe[1], William Burke[1], Andrew Cornish[1], Doug Habberstad[1], Brandon Kopald[2], Alison Bai[2], Yelena Markiv[2], Levi Cole[2], Sara Markuson[2], Yasmine Bensidi-Slimane[2], Alaa Sakallah[2], Katherine Brogan[2], Linnea Lampinen[2], Sara Skiba[2], David Suendermann-Oeft[1], David Pautler[1], Carly Demopoulos[2]

[1]Modality.AI,Inc., [2]University 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.

Acoustic measures	• **Fundamental Frequency (F0):** Minimum value (Hz) and timepoint (s), Maximum value (Hz) and timepoint (s), Mean (Hz), Standard Deviation (Hz) • **Formant Frequency Values:** F1, F2, F3 (Hz) and F2 slope (Hz/s) • **Cepstral Peak Prominence** (CPP in dB) • **Harmonics-to-Noise Ratio** (HNR in dB) • **Articulation time** (in s, excluding pauses) and **speaking time** (in s, including pauses) • **Articulation rate** and **speaking rate** (words per minute) • **Percent pause duration** (%) • **Signal-to-noise ratio** (SNR in dB) • **Articulation intensity** (dB) • **Jitter** and **shimmer** (%)
Visual measures	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

Table 1. Automatically extracted acoustic & visual measures.

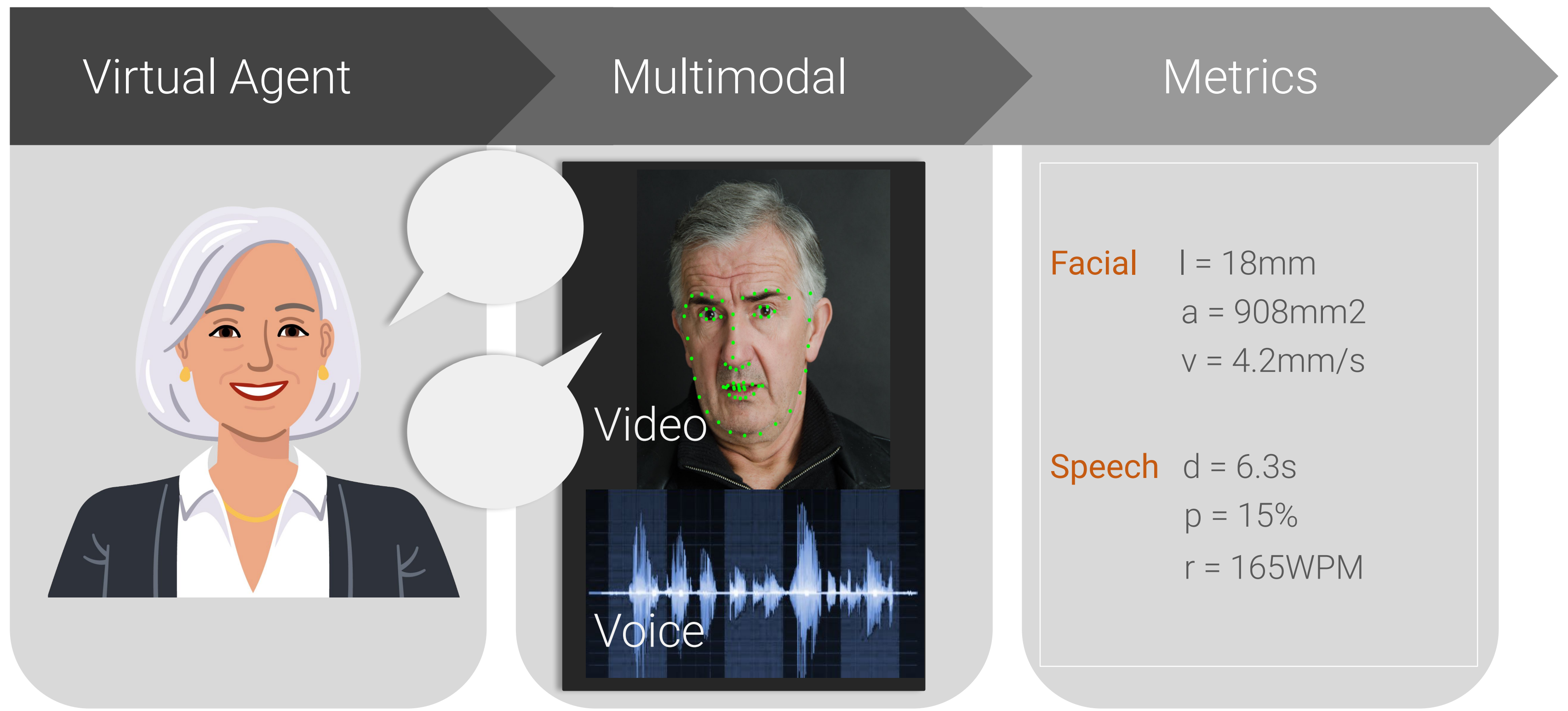

Figure 1. Modality.AI dialogue platform.

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 kinematic 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 **/o/** 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.

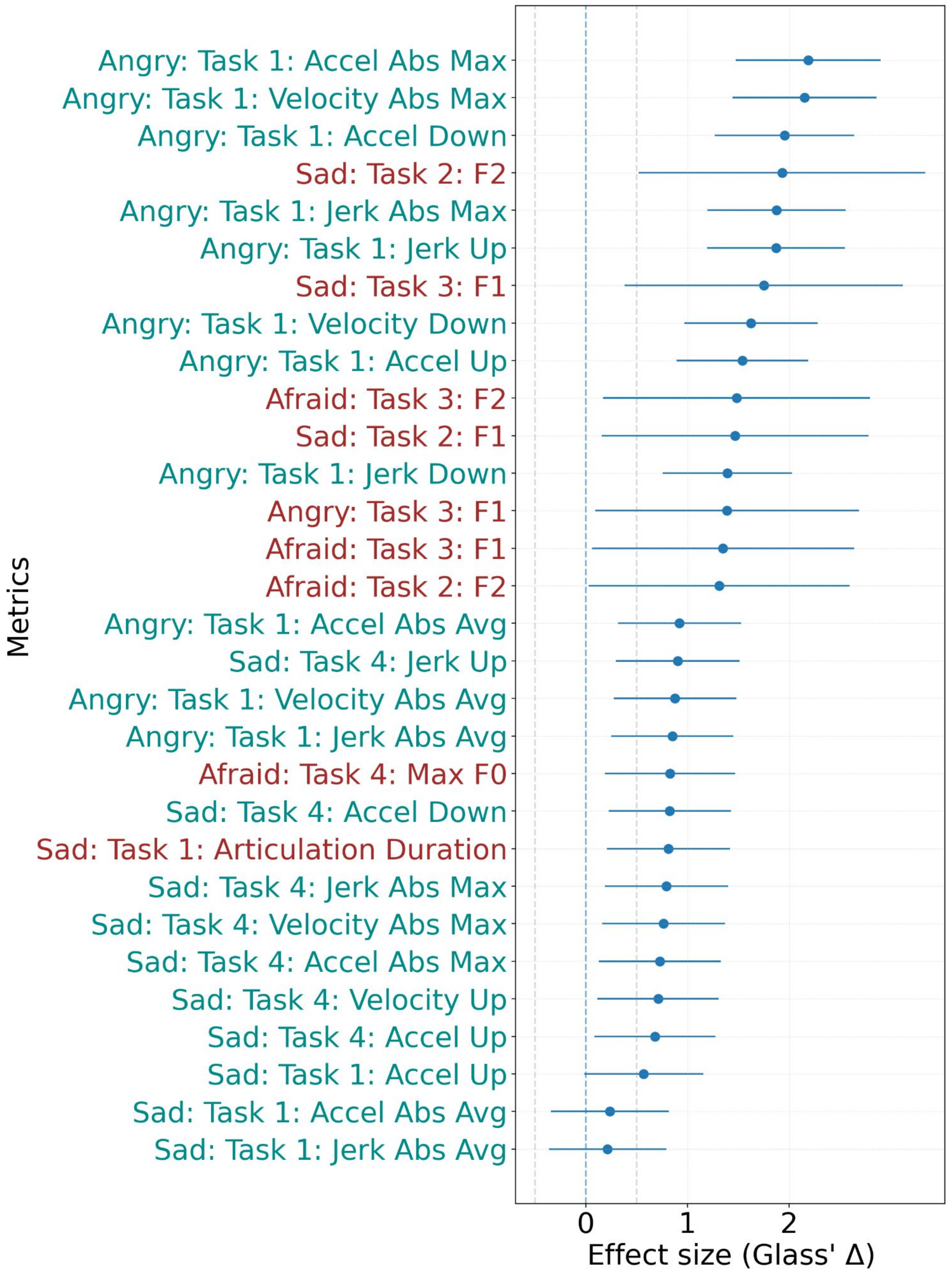

Figure 2. Effect sizes of **acoustic metrics** and **jaw kinematics** that show statistically significant differences between ASD and controls at an alpha threshold of 0.05. Task 1: monosyllable "oh" video stimulus, Task 2: monosyllable "oh" audio stimulus, Task 3: monosyllable "oh" picture stimulus and Task 4: sentence video stimulus.

References

- Hubbard, D. J., et al. (2017). Production and perception of emotional prosody by adults with autism spectrum disorder. Autism Research, 10(12), 1991-2001.
- Loveland, K. A., et al. (1994). Imitation and expression of facial affect in autism. Development and Psychopathology, 6(3), 433-444.
- Kothare, H., et al. (2021) Investigating the Interplay Between Affective, Phonatory and Motoric Subsystems in Autism Spectrum Disorder Using a Multimodal Dialogue Agent. Proc. Interspeech 2021, 1967-1971, doi: 10.21437/Interspeech.2021-1796