Preliminary Investigation of Psychometric Properties of a Novel Multimodal Dialog Based Affect Production Task in Children and Adolescents with Autism

Carly Demopoulos[1], Linnea Lampinen [1], Cristian Preciado [1], Hardik Kothare [2] & Vikram Ramanarayanan[1,2]

[1] University of California, San Francisco
[2] Modality.AI, Inc., San Francisco
carly.demopoulos@ucsf.edu

Abstract

Impairments in nonverbal communication are a defining feature of autism spectrum disorder (ASD) and can manifest as difficulty with, or even complete lack of, communication of emotional states via production of facial affect or vocal affect. The purpose of this study was to evaluate psychometric properties of a novel multimodal dialog based Affect Production Task (APT) in children and adolescents (ages 8-17) with a diagnosis of autism (N=72) or neurotypical controls (N=37). Participants completed activities designed to quantify objective facial and vocal affect production ability using audiovisual capture. Criterion, ecological, and discriminant validity were assessed. Psychometric performance across task conditions, age, sex, and race-ethnicity also was examined. Results of this initial psychometric evaluation suggest that the APT is a valid measure of affect production abilities in children and adolescents, and that psychometric performance is invariant to age, sex, or race/ethnicity.

Index Terms: facial affect, vocal affect, autism, assessment, psychometric properties, Affect Production Task

1. Introduction

1.1. Background

Large scale efforts have been undertaken to address the challenge of precise measurement of social communication skills in autism spectrum disorder (ASD). From these efforts, several limitations were identified in even the best of available measures (e.g., subjectivity, high training burden, inadequate coverage of symptom domains, etc.) [1], [2]. Further, these measures collapse many symptom areas into general domains, whereas, social communication is complex, involving a range of related but distinct skills. As such, overly broad measures of social communication abilities are inadequate for quantifying these skills or being sensitive to their change over time. These limitations have led to a call for development of objective measures of social communication skills [3], [4] to address the unmet assessment needs of researchers and clinicians.

A particular gap in measurement of social communication is in the domain of nonverbal communication. While the literature on language/verbal communication impairment in autism is vast and documents the significant functional impact of deficits in these skills, an important aspect of the success in generating this large body of literature is the availability of objective, performance-based measures of *verbal* communication skills. In contrast, objective, standardized, performance-based measures of *nonverbal* communication are scarce, and typically require laborious processing burden as described in the section to follow. This underscores the need to validate measures of nonverbal communication that can be feasibly employed to sensitively measure specific nonverbal skills in individuals with a range of functional abilities and capture changes in these skills over time. Only then can we begin to understand the impact of these specific deficits and develop empirically supported treatments.

1.2. Prior research on measurement of affect production

Prior affect production research in ASD has largely relied on parent report [5] or rater classification of emotion via a coding system (e.g., Facial Affect Coding System (FACS) [6]–[9], AFFEX [10], Maximally Discriminative Movement Coding System (MAX) [11]) using tasks designed to naturally elicit emotional responses. While parent report measures are an important component of comprehensive assessment, they are vulnerable to the limitations of subjective report, such as underrepresenting measured abilities [12] or capturing nonspecific effects [13], [14]. Likewise, paradigms that spontaneously elicit emotion have methodological limitations (i.e., individual variability in the type and degree of emotional response cannot be controlled for). Thus, these paradigms cannot isolate affect production ability from emotional response. Further, emotion classification coding systems are training and labor intensive and require establishment and maintenance of reliability between raters.

To address these issues, the use of digital phenotyping has been proposed to increase precision of measurement for quantifying a range of behaviors associated with autism, including nonverbal communication [4]. This digital phenotyping technology has been successfully applied to study facial affect production in children with autism via spontaneous emotional response [15], [16] and via a prompted facial affect production task of the Janssen Autism Knowledge Engine (JAKE) research assessment system, which uses FACET automated facial expression analysis software [16]. Because the JAKE assessment system was developed to assess a range of autism symptoms, its depth of assessment in each symptom domain is necessarily limited. Consequently, the JAKE affect production task is unimodal, assessing only facial affect production in isolation, whereas nonverbal communication impairment in ASD may manifest as impaired integration of verbal and nonverbal communication [17]. Likewise, studies of automated classification of vocal affect produced by individuals with ASD have been unimodal and/or have utilized an emotion elicitation design [18], [19], or have limited emotional specificity (i.e., positive, negative, neutral) [20], which is inadequate for capturing the nature of abnormal affect production in this population. In a systematic review of studies

reporting automated emotion recognition systems in autism, several limitations were identified [21], including unimodal assessment of usually facial affect only, small and largely neurotypical samples with few females, use of spontaneous emotion elicitation paradigms which confound individual differences in emotional responsivity, inadequate measures of "ground truth", and reliance on physiological signals for recognition of emotional states, when in fact, physiological signals carry information about emotional intensity, but not emotional classification. We have developed a novel Affect Production Task (APT) to address these limitations in prior approaches to the measurement of affect production in ASD.

In prior work we have demonstrated the predictive validity of the APT in distinguishing between groups of typically developing and autistic youth [22]. Here, we further explore psychometric properties, including criterion, ecological, and discriminant validity, as well as psychometric performance across age, sex, race/ethnicity, and task conditions.

2. Methods

2.1. Participants

Participants were 109 children and adolescents ages 8-17 years who were diagnosed with autism spectrum disorder (ASD; N=72) or were typically developing controls (TDC; N=37). Sex assigned at birth was evenly distributed for the TDC group (18 male, 19 female) and was more balanced than population prevalence for the ASD group (44 male, 28 female).

2.2. Procedures

ASD diagnoses were confirmed according to DSM-5 criteria informed by use of gold standard diagnostic measures, including the Autism Diagnostic Observation Schedule-2nd Edition (ADOS-2) and the Autism Diagnostic Interview-Revised (ADI-R) [23]. Participants were administered the APT as part of a neuropsychological research battery for ongoing studies of speech and voice in autism via a multimodal dialog platform [34], [35] wherein a virtual agent engages participants in a sequence of affect production subtasks as described below.

2.2.1 Affect Production Task

The APT interactive session asks the participants to produce one of four emotions (i.e., happy, sad, angry, afraid) through each of the subtasks described below. The session begins with a speaker test, background noise check and a microphone test. The participant is asked to ensure that their face is fully visible with no face coverings or shadows obscuring their face. The virtual dialogue agent then welcomes the participant to the session and prompts neutral facial and vocal expression 3 times to establish a baseline. There is an option for participants to repeat any incorrect trial (e.g., not ready, unusable response).

Affect production is assessed in 3 task conditions. The first is the Noncontextual Monosyllabic Condition in which the participant is asked to say the word "oh" in a way that communicates the specified emotion (i.e., happy, sad, angry, afraid) by the way their face looks and their voice sounds. The purpose of this task is to assess the ability of the participant to produce a monosyllabic utterance that conveys a specified emotion. The monosyllabic utterance is used to minimize expressive language/speech demands. In the Contextual Monosyllabic Condition, the participant listens to a brief illustrated emotional narrative before being asked to say the word "oh" in a way that communicates the specified emotion by the way their face looks and their voice sounds. The purpose of this task condition is to provide context to help those who may not understand the concept of named emotions in isolation. Finally, in the Noncontextual Sentence Length Condition, the participant is asked to say the emotionally neutral sentence "I'll be right back"[1] in a way that communicates the specified emotion by the way their face looks and their voice sounds.

2.2.2 Computation of Facial and Vocal Metrics

MediaPipe Face Detection based on BlazeFace [25] is used to determine frame-wise x and y coordinates of the face. Facial landmarks are then generated by the Google MediaPipe Face Mesh algorithm [26], 14 of which are key landmarks in the computation of facial movements (Figure 1). All facial metrics derived from these landmarks are measured in pixels and then normalized by dividing them by the inter-caruncular distance (i.e., the distance between the right eye left corner and the left eye right corner; shown in red in Figure 1) to account for cross-participant positional variability relative to the camera [27]. Individual variability in inter-caruncular distance has not been found to impact data analysis in our pilot sample, as outlier analysis has only identified outliers for individual responses (mainly due to noncompliance, distraction, etc.); no outliers were identified on the participant level. Facial metrics are defined in Table 1. Speech data were collected at a sampling rate of 48kHz. Praat [28] was used to extract spectral metrics, energy metrics, and duration metrics (Table 2).

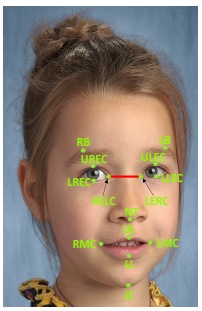

Figure 1: *14 landmarks used to compute facial metrics.* RELC = right eye left corner; LERC = left eye right corner; RB, LB = right/left brow; UREC, ULEC = upper right/left eye center; LREC, LLEC = lower right/left eye center; NT = nose tip; UL = upper lip; LL = lower lip; RMC, LMC = right/left mouth corner; JC = jaw center

2.2.3 Human ratings of affect production ability

Two human raters classified each participant's responses as happy, sad, angry, afraid, or neutral, blinded to the prompted emotion for each response. Facial and vocal responses were classified separately, so that raters only had access to one modality (face or voice) when performing ratings. Human raters had intact affect recognition abilities, determined by scoring in the average range or higher on the Diagnostic Analysis of Nonverbal Accuracy-2 (DANVA-2)[24], a standardized, norm-referenced facial and vocal affect recognition task. While only four emotions were prompted (happy, sad, angry, afraid), raters had a five-response choice option (inclusion of "neutral") to allow them to identify flat affect/failure to convey any emotion. Raters were not instructed on how to classify affect via use of a facial affect coding system or any other operationalized set of

[1] The sentence "I'll be right back" was selected for its emotionally neutral semantic context in an effort to parallel the stimulus used in a standardized facial and vocal affect recognition task, the Diagnostic Analysis of Nonverbal Accuracy-2 (DANVA-2)[24]. The DANVA-2 vocal affect stimuli use the phrase "I'm going out of the room right now, but I'll be back later." We chose a simpler sentence for greater accessibility to those who speak in less complex sentences.

imposed criteria developed based upon other task demands. Instead, we aimed to classify affective expression based on the subjective impression of the rater in order to most closely approximate the experience of one's affect being interpreted in daily life. The agreement between the human rater's classification of affect and the emotion that was prompted was used as an index of ***affect production ability***, as this measures the degree to which the participant was able to effectively communicate the intended emotion via facial/vocal expression. Each human rater's percent accuracy for all responses in each condition was calculated and averaged separately for facial and vocal affect. The averages were then averaged across raters to derive affect production ability scores for each participant.

Table 1. *Glossary of the 12 objective facial metrics*

Metric	Construct
Lip aperture	Average lip opening
Lip width	Average lip width
Mouth surface area (MSA)	Average total MSA
MSA mean symmetry ratio	Symmetry of the mouth
Velocity, acceleration & jerk of lower lip & jaw center	Speed & rates of change in speed & acceleration
Eye opening	Average opening of eyes
Vertical displacement of the eyebrows (VDE)	Average vertical eyebrow displacement

Table 2. *Glossary of the 17 objective vocal metrics*

Metric	Construct
Shimmer [%]	Perturbation of the amplitude domain
Signal-to-noise ratio [dB]	Ratio of signal power to noise power
Intensity [dB]	Energy of a sound over an area
Articulation intensity [dB]	Intensity of speech in an utterance
Fundamental frequency [Hz]	Vocal pitch, with typical ranges of values for different sexes and ages
Jitter [%]	Frequency variation from cycle to cycle
Formant frequencies [Hz] (F1, F2, F3)	Distinctive frequency components of acoustic signal produced by speech; characterizes vowel quality
F2 slope [Hz/s]	Rate of vocal tract shape change for vowels
Speaking duration [s]	Total duration of the utterance
Pause time [%]	% of time utterance is paused
Articulation duration [s]	Amount of voiced time
Timepoint [s] of max & min F0	Latency of maximum and minimum pitch across the utterance
Cepstral peak prominence [dB]	Robust measure of voice quality; lower values indicate greater levels of dysphonia
Harmonics-to-noise ratio [dB]	Perceived hoarseness, breathiness or roughness (lower = more hoarse)

2.3. Analysis of validity

2.3.1 Criterion-related validity

Criterion-related validity was assessed via the prediction of facial and vocal affect production ability from the objective facial and vocal metrics, respectively. The combined participant sample was used for these analyses in order to capture a range of impaired and intact affect production abilities to be compared against objective metrics. Stepwise linear regression analyses were performed for each task condition to examine the prediction of affect production ability from the objective automated metrics. All analyses were performed separately for facial and vocal affect production. Participants were excluded from these analyses if missing data exceeded 20% for objective metrics (e.g., missing data for entire task conditions). For those included, missing values were imputed with mean replacement.

2.3.2 Ecological validity

Ecological validity was assessed via nonparametric correlation analyses between affect production ability and clinician rating of ASD symptom severity on the Childhood Autism Rating Scale-2nd Edition (CARS-2) [29], as well as parent report of facial and vocal expression on the ADI-R [30]. These analyses were performed with the ASD group only as the TDC group was not evaluated on these autism diagnostic tools.

2.3.3 Discriminant validity

Discriminant validity was assessed via hierarchical linear regression analyses to quantify the relative contribution of objective metrics to the prediction of human rated affect production ability after controlling for performance on measures of facial and vocal sensorimotor control and affect recognition. Sensorimotor control of face and voice was assessed via rater accuracy for imitated facial expressions (i.e., if a rater identified an expression as happy, and the participant was imitating the actor's portrayal of happy, this indicates adequate sensorimotor imitation). Facial and vocal affect recognition were indexed by averaging the age-scaled scores for adult and child conditions of the DANVA-2 faces and paralanguage subtests, respectively [24]. Associations also were examined between affect production ability and nonsocial ASD characteristics as rated by parents on the RBS-R [31], [32] using the five-factor scoring [33]. Analyses were performed separately for face and voice for all analyses. Discriminant validity analyses were performed on the ASD sample to avoid inflation of measures of association due to group effects.

2.4. Psychometric properties across sex, age, and race/ethnicity, and task demands

Analyses for validity measures were performed controlling for age and separately for participants assigned male versus female sex at birth and for participants in different racial/ethnic groups to evaluate whether APT performance was invariant to age, sex, and race/ethnicity. Repeated measures ANOVA was performed across all participants to examine within-subject effects of task condition to determine if different task demands resulted in different affect production ability scores within-individual or interacted with group membership.

3. Results

3.1. Criterion validity

Objective automated facial and vocal metrics significantly predicted affect production ability. Specifically, stepwise linear regression analyses indicated that facial metrics predicted 60% of the variance in rater accuracy for both noncontextual production tasks, and 32% of the variance for the contextual production task. Objective vocal metrics predicted 41% of the variance in rater accuracy for the noncontextual monosyllabic production task, and 58% of the variance for the noncontextual sentence-length task and the contextual production task.

3.2. Ecological validity

Correlation analyses indicate that clinician ratings of higher overall ASD symptom severity on the CARS-2 was

significantly associated with lower affect production ability, quantified by accuracy of human rater classifications for facial (r=-.391, p=.013, N=40) and vocal production (r=-.434, p=.007, N=38). Scatterplots also show correspondence between APT performance and parent ratings of facial and vocal expression on the ADI-R. The distribution of ratings shows a trend toward lower rater accuracy corresponding to higher (more severe) ADI-R rating (Figure 2).

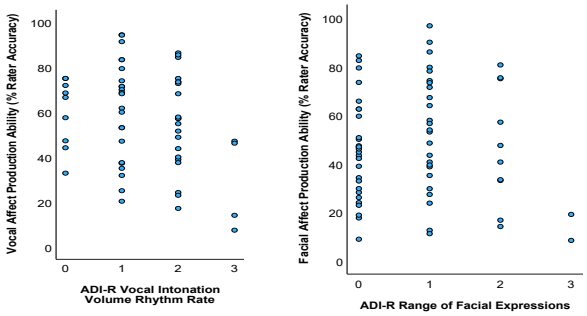

Figure 2: *Scatterplots of Affect Production Ability and Parent Report of Facial and Vocal Expression*

3.3. Discriminant validity

The linear combination of raw APT metrics contributed additional, statistically significant proportions of the variance in human rater accuracy for all production tasks, after controlling for affect recognition and sensorimotor control. Specifically, APT metrics contributed an additional 33% variance for facial monosyllabic production, 27% for facial sentence-length, 42% for facial contextual monosyllabic, 57% for vocal monosyllabic production, 72% for vocal sentence-length, 30% for vocal contextual monosyllabic. Correlation coefficients were low across all RBS-R scales and facial and vocal affect production ability scores (≤ .203).

3.4. Psychometric properties across sex, age, and race/ethnicity, and task demands

Associations between CARS-2 scores and affect production ability for facial (-.438, p=.005) and vocal affect (r=-.593, p<.001) remained strong after controlling for age. With regard to effects of sex, associations between CARS-2 and APT facial and vocal affect production were run separately for participants assigned male versus female sex at birth. These associations remained strong in the split sample, with r=-.436 (facial affect; N=22) and r=-.394 (vocal affect; N=21) for male participants. Likewise, for females, r=-.313 for facial affect (N=18) and r=-.485 for vocal affect (N=17). ANOVA results indicate that neither facial nor vocal affect production differed significantly between racial or ethnic groups. Analyses were not sufficiently powered to examine interaction effects of diagnostic group on differential impact of race or ethnicity on task performance. Finally, repeated measures ANOVA indicated that, for facial affect production, there were no significant interaction effects of group and task condition; however, the ASD group performed significantly more poorly than the TDC group across all conditions. For vocal affect production, there was a significant effect of task (F(2,88)=4.184, p=.018), but the interaction between group and task was not statistically significant. Post hoc contrasts indicated that the TDC group performed significantly better for the Contextual Monosyllabic Condition relative to their performance in the other conditions.

4. Discussion and Conclusions

This study assessed psychometric properties of the APT, a novel automated measure of affect production. Building on prior work demonstrating predictive validity of the APT automated metrics in classifying ASD versus TDC groups [22], this study investigated the criterion, ecological, and discriminant validity of the APT, as well as the effects of age, sex, race, ethnicity and task demands on psychometric performance. Investigation of criterion validity revealed that the automated facial and vocal metrics captured during the APT predicted the effectiveness of the participant's affect production (i.e., how well the intended emotion was communicated to human raters). This successful prediction was demonstrated in the absence of any weighting or other optimization of metrics, suggesting that machine learning can be applied to develop an APT affect recognition algorithm from a large sample of APT data. This would allow for automation of human ratings, which would save time and allow for standardization and collection of normative data on APT performance.

Assessment of ecological and discriminant validity indicated that APT performance was significantly associated with severity of overall ASD symptoms on the CARS-2, but was not associated with nonsocial ASD symptoms measured on the RBS-R. Further, scatterplots of APT performance and parent reported facial and vocal expression on the ADI-R indicated that when parents rated these as more abnormal, affect production ability scores were lowest. Further, objective metrics successfully predicted affect production ability even after controlling for sensorimotor control of face and voice and affect recognition abilities. Taken together, these findings suggest that the APT captures skills in affective communication above and beyond what can be attributed to basic sensorimotor control or emotional understanding more broadly.

Finally, evidence of ecological validity was preserved after covarying age and across sexes. Affect production abilities did not differ across racial and ethnic groups. Likewise, for facial affect production, the TDC group outperformed the ASD group across all task conditions, but group differences did not significantly differ across tasks. In contrast, for vocal affect production there was a significant effect of task on performance; however, this was driven mainly by the TDC group showing an advantage on the contextual as opposed to noncontextual tasks, whereas the ASD group did not show the same benefit from provision of narrative context in producing affective expressions. These findings suggest that ability to incorporate situational context into affect production may explain some of the difficulty in nonverbal communication that is a defining feature of autism. Notably, sample size and inability to include TDC participants in analyses involving ASD diagnostic measures limit the extent of the analyses we were able to perform. Future research in large sample of individuals with and without ASD a broader age range are necessary to establish validity and reliability of the APT.

In summary, preliminary evaluation of the psychometric properties suggests that the standardized task structure of the APT is effective in capturing affect production abilities in children and adolescents across age, sex, race, ethnicity, and task demands. This suggests that the APT is accessible to individuals of different demographic and verbal abilities. Further, unweighted automated metrics of facial and vocal features successfully predicted affect production abilities, suggesting potential for further automation and standardization of the APT.

5. References

[1] S. Faja *et al.*, "Evaluation of clinical assessments of social abilities for use in autism clinical trials by the autism biomarkers consortium for clinical trials," *Autism Res.*, vol. 16, no. 5, pp. 981–996, 2023, doi: 10.1002/aur.2905.

[2] E. Anagnostou *et al.*, "Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder," *Autism*, vol. 19, no. 5, pp. 622–636, 2015, doi: 10.1177/1362361314542955.

[3] B. P. Wang, "The need for objective measures to intervention research in autism," *SFARI Blog*, pp. 1–8, 2019.

[4] G. Dawson, "Assessment of outcomes in autism clinical trials over the course of development," *Eur. Neuropsychopharmacol.*, vol. 48, pp. 40–41, 2021, doi: 10.1016/j.euroneuro.2021.02.018.

[5] L. Capps, C. Kasari, N. Yirmiya, and M. Sigman, "Parental Perception of Emotional Expressiveness in Children With Autism," *J. Consult. Clin. Psychol.*, vol. 61, no. 3, pp. 475–484, 1993, doi: 10.1037/0022-006X.61.3.475.

[6] E. M. Weiss, C. Rominger, E. Hofer, A. Fink, and I. Papousek, "Less differentiated facial responses to naturalistic films of another person's emotional expressions in adolescents and adults with High-Functioning Autism Spectrum Disorder," *Prog. Neuro-Psychopharmacology Biol. Psychiatry*, vol. 89, no. August 2018, pp. 341–346, 2019, doi: 10.1016/j.pnpbp.2018.10.007.

[7] J. Legiša, D. S. Messinger, E. Kermol, and L. Marlier, "Emotional responses to odors in children with high-functioning autism: Autonomic arousal, facial behavior and self-report," *J. Autism Dev. Disord.*, vol. 43, no. 4, pp. 869–879, 2013, doi: 10.1007/s10803-012-1629-2.

[8] T. Denmark, J. Atkinson, R. Campbell, and J. Swettenham, "Signing with the Face: Emotional Expression in Narrative Production in Deaf Children with Autism Spectrum Disorder," *J. Autism Dev. Disord.*, vol. 49, no. 1, pp. 294–306, 2019, doi: 10.1007/s10803-018-3756-x.

[9] M. E. Snow, M. E. Hertzig, and T. Shapiro, "Expression of Emotion in Young Autistic Children," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 26, no. 6, pp. 836–838, 1987, doi: 10.1097/00004583-198726060-00006.

[10] S. Macari *et al.*, "Emotional Expressivity in Toddlers With Autism Spectrum Disorder," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 57, no. 11, pp. 828-836.e2, 2018, doi: 10.1016/j.jaac.2018.07.872.

[11] C. Kasari, M. Sigman, P. Mundy, and N. Yirmiya, "Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children," *J. Autism Dev. Disord.*, vol. 20, no. 1, pp. 87–100, 1990, doi: 10.1007/BF02206859.

[12] S. James, S. Hallur, J. Anbar, N. Matthews, K. Pierce, and C. J. Smith, "Consistency between parent report and direct assessment of development in toddlers with autism spectrum disorder and other delays: Does sex assigned at birth matter?," *Autism Res.*, vol. 16, no. 6, pp. 1174–1184, 2023, doi: 10.1002/aur.2927.

[13] S. K. Bennetts, F. K. Mensah, E. M. Westrupp, N. J. Hackworth, and S. Reilly, "The agreement between parent-reported and directly measured child language and parenting behaviors," *Front. Psychol.*, vol. 7, no. NOV, 2016, doi: 10.3389/fpsyg.2016.01710.

[14] K. D. Ten Eycke and D. Dewey, "Parent-report and performance-based measures of executive function assess different constructs," *Child Neuropsychol.*, vol. 22, no. 8, pp. 889–906, 2016, doi: 10.1080/09297049.2015.1065961.

[15] H. L. Egger *et al.*, "Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study," *npj Digit. Med.*, vol. 1, no. 1, 2018, doi: 10.1038/s41746-018-0024-6.

[16] J. Manfredonia *et al.*, "Automatic Recognition of Posed Facial Expression of Emotion in Individuals with Autism Spectrum Disorder," *J. Autism Dev. Disord.*, vol. 49, no. 1, pp. 279–293, 2019, doi: 10.1007/s10803-018-3757-9.

[17] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule-Second Edition (ADOS-2)." Western Psychological Services, Torrance, CA, 2012.

[18] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, and D. Cohen, "Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children Chair of Complex & Intelligent Systems , University of Passau , Germany Institute of Intelligent Systems and Robotics , Universit ´," *Interspeech*, pp. 1210–1214, 2016.

[19] P. Sukumaran and K. Govardhanan, "Towards Voice Based Prediction and Analysis of Emotions in ASD Children," *J. Intell. Fuzzy Syst.*, vol. 41, no. 5, pp. 5317–5326, 2021.

[20] J. Li, A. Bhat, and R. Barmaki, "A two-stage multi-modal affect analysis framework for children with autism spectrum disorder," *CEUR Workshop Proc.*, vol. 2897, pp. 7–14, 2021.

[21] A. Landowska *et al.*, "Automatic Emotion Recognition in Children with Autism: A Systematic Literature Review," *Sensors*, vol. 22, no. 4, pp. 1–29, 2022, doi: 10.3390/s22041649.

[22] H. Kothare *et al.*, "Atypical speech acoustics and jaw kinematics during affect production in children with Autism Spectrum Disorder assessed by an interactive multimodal conversational platform," 2022.

[23] S. L. Bishop and C. Lord, "Commentary: Best practices and processes for assessment of autism spectrum disorder – the intended role of standardized diagnostic instruments," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 64, no. 5, pp. 834–838, 2023, doi: 10.1111/jcpp.13802.

[24] S. Nowicki, "Manual for the receptive tests of the Diagnostic Analysis of Nonverbal Accuracy 2." Atlanta ,GA, 2010.

[25] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. & Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *arXiv Prepr. arXiv1907.05047*, 2019.

[26] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *arXiv Prepr. arXiv1907.06724*, 2019.

[27] O. Roesler *et al.*, "Exploring Facial Metric Normalization For Within- and Between-Subject Comparisons in a Multimodal Health Monitoring Agent," 2022, doi: https://doi.org/10.1145/3536220.3558071.

[28] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT," *Glot Int.*, vol. 5, no. 9, pp. 341–347, 2001.

[29] E. Schopler, M. E. Van Bourgondien, G. J. Wellman, and S. Love, "Childhood Autism Rating Scale, Second Edition (CARS-2)." Western Psychological Services, Torrance, CA, 2010.

[30] C. Lord, M. Rutter, and A. Le Couteur, "Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders.," *J. Autism Dev. Disord.*, vol. 24, no. 5, pp. 659–685, Oct. 1994.

[31] J. Bodfish, F. Symons, and M. Lewis, "The Repetitive Behavior Scale." Western Carolina Center Research Reports, 1999.

[32] J. Bodfish and M. Lewis, "Repetitive Behavior in Autism," 2002.

[33] K. S. L. Lam and M. G. Aman, "The repetitive behavior scale-revised: Independent validation in individuals with autism spectrum disorders," *J. Autism Dev. Disord.*, vol. 37, no. 5, pp. 855–866, 2007, doi: 10.1007/s10803-006-0213-z.

[34] V. Ramanarayanan *et al.*, "When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do," in Proc. INTERSPEECH 2023, 2023, pp. 678–679.

[35] V. Ramanarayanan, "Multimodal technologies for remote assessment of neurological and mental health," Journal of Speech, Language, and Hearing Research, pp. 1–8, 2024.