

1 Reliably quantifying the severity of social symptoms in 2 children with autism using ASDSpeech

3 Marina Eni^{1,3}✉, Michal Ilan^{2,3}, Analya Michaelovski^{3,4}, Hava M. Golan^{5,3}, Gal Meiri^{2,3}, Idan
4 Menashe^{6,3}, Ilan Dinstein^{7,8,3}, and Yaniv Zigel^{1,3}

- 5 1. Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva,
6 Israel.
- 7 2. Pre-School Psychiatry Unit, Soroka University Medical Center, Beer-Sheva, Israel.
- 8 3. Azrieli National Centre for Autism and Neurodevelopment Research, Ben-Gurion University
9 of the Negev, Beer-Sheva, Israel.
- 10 4. Zusman Child Development Center, Soroka University Medical Center, Israel.
- 11 5. Department of Physiology and Cell Biology, Faculty of Health Sciences, Ben-Gurion
12 University of the Negev, Beer-Sheva, Israel.
- 13 6. Department of Public Health, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
- 14 7. Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
- 15 8. Department of Brain & Cognitive Sciences, Ben-Gurion University of the Negev, Beer-Sheva,
16 Israel.

17 ✉email: marinamu@post.bgu.ac.il.

19 **Abstract**

20 Several studies have demonstrated that the severity of social communication problems, a core
21 symptom of Autism Spectrum Disorder (ASD), is correlated with specific speech characteristics
22 of ASD individuals. This suggests that it may be possible to develop speech analysis algorithms
23 that can quantify ASD symptom severity from speech recordings in a direct and objective manner.
24 Here we demonstrate the utility of a new open-source AI algorithm, ASDSpeech, which can
25 analyze speech recordings of ASD children and reliably quantify their social communication
26 difficulties across multiple developmental timepoints. The algorithm was trained and tested on the
27 largest ASD speech dataset available to date, which contained 99,193 vocalizations from 197 ASD
28 children recorded in 258 Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2)
29 assessments. ASDSpeech was trained with acoustic and conversational features extracted from the
30 speech recordings of 136 children, who participated in a single ADOS-2 assessment, and tested
31 with independent recordings of 61 additional children who completed two ADOS-2 assessments,

32 separated by 1–2 years. Estimated total ADOS-2 scores in the test set were significantly correlated
33 with actual scores when examining either the first ($r(59) = 0.544$, $P < 0.0001$) or second ($r(59) =$
34 0.605 , $P < 0.0001$) assessment. Separate estimation of social communication and restricted and
35 repetitive behavior symptoms revealed that ASDSpeech was particularly accurate at estimating
36 social communication symptoms (i.e., ADOS-2 social affect scores). These results demonstrate
37 the potential utility of ASDSpeech for enhancing basic and clinical ASD research as well as
38 clinical management. We openly share both algorithm and speech feature dataset for use and
39 further development by the community.

40 **Introduction**

41 Autism Spectrum Disorder (ASD) is diagnosed by the presence of social communication
42 difficulties and the existence of Restricted and Repetitive Behaviors (RRBs)¹. Most ASD children
43 exhibit language delays during early childhood², with 25–30% remaining minimally verbal (i.e.,
44 use < 50 words) throughout childhood³. However, core ASD symptoms are not necessarily evident
45 in the amount of speech produced by an individual and may instead be evident in the way they
46 speak. Some ASD children exhibit poorer fluency⁴, echolalia (i.e., speech repetition)⁵, mix
47 pronouns⁶, and use atypical articulation and prosody^{7,8} that are apparent in the acoustic features of
48 their vocalizations^{9,10}. Studies have reported, for example, that verbal ASD children tend to speak
49 with higher pitch and larger pitch variability than typically developing (TD) children^{8,9}. ASD
50 children also exhibit significantly fewer phoneme vocalizations¹¹, fewer conversational turns (i.e.,
51 reciprocating in a conversation)^{11–13}, more non-speech vocalizations^{12,14}, more distressed
52 vocalizations (crying, screaming)¹⁵, and a lower ratio of syllables to vocalizations¹⁶ than TD
53 children.

54 Several studies have used automated speech analysis techniques to classify ASD and TD
55 children based on extracted speech features^{17–24}. In some studies, diagnostic classification was
56 based on linguistic features such as vocabulary and fluency²⁴ while in others it was based on
57 acoustic features such as pitch^{18–20,22,23}, jitter^{20,23}, shimmer^{20,23}, energy^{18,19}, Zero-Crossing Rate
58 (ZCR)^{18,19}, and Mel-Frequency Cepstral Coefficients (MFCCs)¹⁹.

59 Three recent studies have extended this research by training machine and deep learning
60 algorithms to estimate ASD severity according to extracted speech features. In all these studies
61 ground truth was established by clinicians using the Autism Diagnostic Observation Schedule 2nd
62 edition (ADOS-2), a semi-structured assessment where clinicians score the behavior of children
63 during specific tasks/games²⁵. The ADOS-2 yields a total severity score as well as separate Social
64 Affect (SA) and Restricted and Repetitive Behaviors (RRB) scores that quantify social difficulties
65 and RRB symptoms, respectively. In the first study, the authors extracted vocalization rates and
66 durations from speech recordings of 33 ASD children during an ADOS-2 assessment and reported
67 that a trained synthetic random forest model was able to accurately estimate their ADOS-2 Social

68 Affect (SA) scores²⁶. Another study extracted hundreds of conversational, acoustic, and lexical
69 speech features from speech recordings of 88 adolescents and adults with ASD during an ADOS
70 assessment (1st edition) and reported that a trained Deep Neural Network (DNN) was able to
71 accurately estimate scores of four specific ADOS items that quantify the ability to maintain a
72 mature social conversation²⁷. Finally, in the third study, from our group, we extracted acoustic
73 features such as pitch and energy, and conversational features such as turn-taking and speech rate
74 from speech recordings of 72 children (56 with ASD) during an ADOS-2 assessment²⁸. We
75 demonstrated that a trained Convolutional Neural Network (CNN) model was able to accurately
76 estimate total ADOS-2 scores across multiple train-test subsamples.

77 While these results are encouraging, algorithms developed so far were trained and tested with
78 relatively small ASD samples that are not likely to represent the large heterogeneity of speech
79 styles and characteristics in the broad ASD population²⁹. Moreover, previous studies examined
80 only a single timepoint of data from each participant, thereby limiting the ability to assess the
81 reliability of algorithms to assess ASD symptom severity at different developmental timepoints.
82 Previous studies also did not compare the ability of deep learning models to successfully estimate
83 the severity of social ASD symptoms versus RRB symptoms. Most importantly, previous studies
84 did not share their algorithms and data in a transparent manner that would enable re-production of
85 results and further development of algorithms by the research community.

86 To address these limitations, we created the largest speech recording dataset available to date,
87 which contained 99,193 vocalizations from 197 ASD children recorded in 258 ADOS-2
88 assessments, with 61 of the children participating in two ADOS-2 assessments that were separated
89 by 1-2 years. This comprehensive dataset enabled us to train and test the ASDSpeech algorithm
90 on different subsets of children and compare its accuracy across two developmental timepoints as
91 well as sex and age sub-groups. In addition, we also examined the ability to estimate ADOS-2 SA
92 versus RRB scores (i.e., social difficulties versus RRB symptoms). We intentionally used raw
93 ADOS-2 scores, which have a considerably wider range than ADOS-2 calibrated severity scores
94 ^{30,31}, thereby increasing the potential sensitivity of the algorithm. Finally, we openly share the
95 algorithm and speech feature dataset to promote transparency and enable further use and
96 development by the research community.

97 **Methods**

98 **Participants and setting.** We analyzed data collected at the Azrieli National Centre for Autism
99 and Neurodevelopment Research (ANCAN), a collaboration between Ben-Gurion University of
100 the Negev (BGU) and eight partner clinical centers where ASD is diagnosed throughout Israel.
101 ANCAN manages the national autism database of Israel with data from > 3000 children in 2023
102 and growing^{32,33}. All recordings used in the current study were performed in a single ANCAN
103 assessment room located at Soroka University Medical Center (SUMC), the largest partner clinical

104 site. A total of 197 children (1–7-years-old) who completed at least one ADOS-2 assessment
 105 between 2015 and 2021 and received an ASD diagnosis were included in this study (Table 1). Of
 106 the participating children, 136 completed a single ADOS-2 assessment and 61 completed two
 107 ADOS-2 assessments at two timepoints separated by 10–29 months, yielding 258 ADOS-2
 108 assessments in total. All ADOS-2 assessments were performed by a clinician with research
 109 reliability. In addition, all participating children had ASD diagnoses that were confirmed by both
 110 a developmental psychologist and either a child psychiatrist or a pediatric neurologist, according
 111 to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria³⁴.
 112 Informed consent was obtained from all parents, and the study received approval from the SUMC
 113 Helsinki committee.

114 **Table 1.** Participating children’s characteristics

	Single assessment	Two assessments	
		T1	T2
	(N = 136)	(N = 61)	(N = 61)
Mean (SD)			
Age (years)	4.26 (1.34)	3.67 (0.98)	4.95 (0.94)
ADOS-2 Total	14.99 (5.88)	14.92 (5.85)	14.92 (5.61)
ADOS-2 SA	10.75 (4.96)	11.26 (5.14)	10.57 (4.55)
ADOS-2 RRB	4.24 (1.88)	3.66 (1.66)	4.34 (1.87)
N (%)			
Sex			
Male	108 (79)	42 (69)	42 (69)
Female	28 (21)	19 (31)	19 (31)
Module			
Module T	17 (13)	9 (15)	0 (0)
Module 1	45 (33)	329 (48)	28 (46)
Module 2	40 (29)	21 (34)	19 (31)
Module 3	3 (25)	2 (3)	14 (23)

115

116 **ADOS-2 assessments.** ADOS-2 is a semi-structured behavioral assessment where a clinician
117 administers specific tasks, observes the behavior of the child, and scores their behavior³⁵. The total
118 ADOS-2 score (range: 0–30) is the sum of the Social Affect (SA, range: 0–22) and Restricted and
119 Repetitive Behavior (RRB, range: 0–8) scores, with higher scores indicating more severe
120 symptoms.

121 **Recording setup.** All ADOS-2 recordings were performed using a single microphone (CHM99,
122 AKG, Vienna) located on a wall, ~1–2m from the child, and connected to a sound card (US-16x08,
123 TASCAM, California). Each ADOS-2 session lasted ~40-minutes (40.75 ± 11.95 min) and was
124 recorded at a sampling rate of 44.1 kHz, 16 bits/sample (down-sampled to 16 kHz).

125 **Detection of child vocalizations.** We manually labeled segments with child vocalizations in each
126 of the audio recordings. These segments included speech, laughing, moaning, crying, and
127 screaming. The child segments often contained multiple vocalizations (e.g., multiple utterances)
128 separated by silence. We separated each segment into multiple vocalizations using energy
129 thresholds of 2.79dB and 0.4dB above the background noise to define the beginning and end of
130 each vocalization, respectively²⁸ (Supplementary Figure S1). Vocalizations that were shorter than
131 110ms were excluded from further analysis (too short to contain an utterance).

132 **Features.** We extracted 49 speech features from the child vocalizations that were categorized into
133 nine groups: pitch, formants, jitter, voicing, energy, Zero-Crossing Rate (ZCR), spectral slope,
134 duration, and quantity/number of vocalizations. All features, except duration and quantity, were
135 first extracted in 40ms windows (window overlap of 75%), resulting in a vector of feature values
136 per vocalization. The minimum, the maximum, and the mean pitch of the voiced vocalizations
137 (across windows) were computed, deriving one value for each vocalization. We then selected a
138 group of 10 consecutive vocalizations and computed the mean and variance across vocalizations
139 for relevant features (Supplementary Table S1). We also computed the mean duration of
140 vocalizations and the overall number of vocalizations in the recording. Taken together, these steps
141 yielded a vector with 49 values corresponding to the 49 features per 10 vocalizations. We
142 performed this procedure 100 times, selecting random groups of ten consecutive vocalizations
143 from the recording. Combining these 100 samples yielded a features matrix of 100×49 per child
144 (Supplementary Figure S2), with the last column (quantity of vocalizations) containing the same
145 value across all rows. Features included:

146

147 ***Frequency related features:***

- 148 • *Pitch (F0)*: Vocal cords vibration frequency (the fundamental frequency) that exists only
149 in voiced speech (e.g., vowels). Voiced Vocalization (VV) was defined as a vocalization
150 where most of its frames ($\geq 60\%$)¹⁰ were voiced (voicing threshold 0.45).
- 151 • *Formants*: The resonant frequencies of the vocal tract that shape vowel sounds³⁶. The first
152 two formants (F1 and F2) relate to tongue position (vertical and horizontal) and influence
153 vowel quality. Their bandwidths affect the clarity of speech.
- 154 • *Jitter*: Variation across adjacent pitch values representing frequency instability³⁷.
- 155 • *Voicing*: Pitch peak amplitude as determined by the autocorrelation function.

156 Pitch and formants were calculated using the PRAAT software³⁸, with a pitch range set to 60–1600
157 Hz (a wide range to increase sensitivity to atypical vocal characteristics).

158 ***Energy/amplitude related features:***

- 159 • *Energy*: The energy ratio between each child’s vocalization and the background noise. The
160 background noise energy was calculated from the energy values extracted from the lowest
161 5% of the recording’s frames.

162 ***Spectral features:***

- 163 • *Zero-Crossing Rate (ZCR)*: The number of zero-crossings apparent in audio segments with
164 child vocalizations³⁹.
- 165 • *Spectral slope*: The slope of the linear regression on the logarithmic power spectrum within
166 the frequency bands of 20–500 Hz (lower band) and 500–1500 Hz (higher band)^{40,41}.

167 ***Conversational features:***

- 168 • *Duration*: Child’s mean vocalization length.
- 169 • *Quantity*: The total number of vocalizations.

170 All features, except for Pitch and Formants, were extracted with custom-written code in Matlab
171 (Mathworks, Inc.).

172 **Training and testing ASDSpeech.** Training was performed with data from the 136 children
173 who completed a single ADOS-2 session only. Feature matrices were used to train two deep
174 learning models with an identical CNN architecture (Supplementary Figure S3). The first model
175 estimated ADOS-2 SA scores and the second estimated ADOS-2 RRB scores. Training was based
176 on minimizing the Mean Squares Error (MSE) of a regression analysis between estimated and

177 actual scores, using the RMSprop (Root Mean Square Propagation) as the optimization
178 algorithm⁴². The training process was preformed 25 times, creating 25 different SA and 25 RRB
179 models that were trained with different combinations of training data sub-samples and learning
180 parameters. We considered this analogous to having 25 clinicians, each with a different learning
181 style and different clinical experience. First, we performed the feature extraction procedure
182 described above 5 times for each child. Since feature extraction included a random selection of
183 consecutive vocalizations, this resulted in 5 different sub-samples of the data. When training each
184 model (separately for SA and RRB) we split the training data into a training-set (80%) and
185 validation set (20%) and applied a random search algorithm to optimize the following learning
186 parameters: batch size, number of epochs, and learning rate, while applying early stopping of
187 patience after 20 epochs to reduce overfitting⁴³. Optimal learning hyper-parameters were selected
188 based on the highest concordance correlation coefficient⁴⁴, between estimated and actual ADOS-
189 2 scores in the training and validation sets respectively. This procedure was performed 5 times
190 using different selections of validation data (i.e., 5-fold cross validation), yielding 5 models with
191 different learning parameters per data sub-sample and 25 models in total for SA and RRB scores
192 separately.

193 Testing was performed with an entirely independent dataset of 61 ASD children who
194 completed two ADOS-2 assessments. For each of these children we estimated a separate SA and
195 RRB score from each of the 25 models described above and then computed their mean, yielding a
196 single SA and RRB score per child. This is analogous to a clinical consensus across the 25 models.
197 Accuracy of ASDSpeech estimation was measured using Pearson correlation and NRMSE (RMSE
198 / $(y_{\max} - y_{\min})$, where y is the actual ADOS-2 score), which were calculated between the estimated
199 and actual ADOS-2 scores in the testing dataset, separately for the first and second ADOS-2
200 assessments (i.e., T1 and T2).

201 **Hardware.** All model training, optimization, and training were performed using custom-written
202 code in Python 3.9.13 using a Keras API 2.6.0 with TensorFlow (version 2.6.0) backend. The
203 training was conducted on an Intel® Xeon® Gold 6140 CPU @ 2.30GHz and NVIDIA GPU Tesla
204 T4.

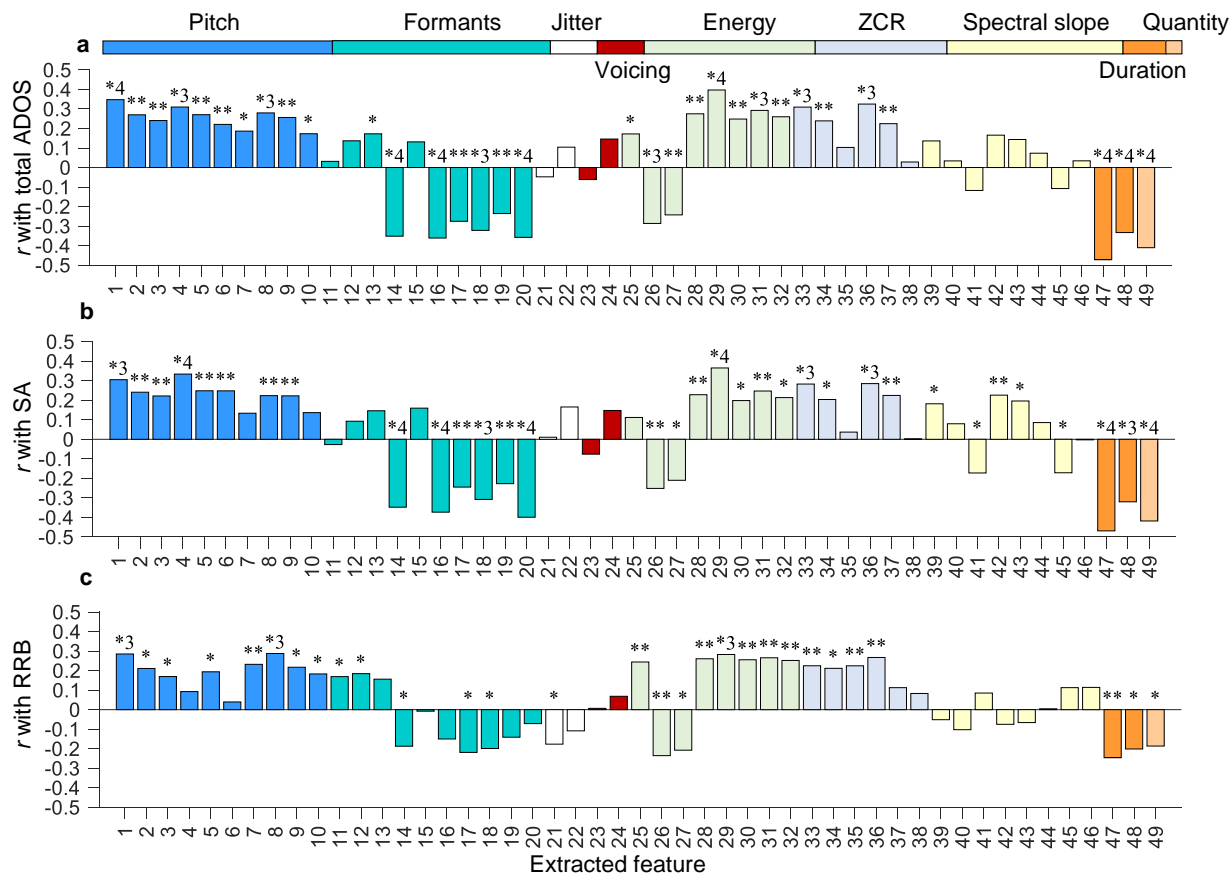
205 **Statistical Analysis.** All statistical analyses were conducted using custom-written code in
206 Python. Associations between speech features and ADOS-2 scores were assessed using Pearson
207 correlations. To evaluate their statistical significance, we performed a random permutation test. In
208 this test, we randomly shuffled the actual ADOS-2 scores across children and calculated the
209 correlation between each feature and the shuffled scores. This randomization procedure was
210 performed 1,000 times, generating a null distribution of random correlation values as computed
211 from the original data that is not necessarily normally distributed as assumed by parametric
212 statistical tests. For a correlation between a speech feature and ADOS-2 score to be considered
213 significant, the actual correlation value had to be higher than the 97.5 percentile of the null

214 distribution. We used an equivalent analysis to assess the statistical significance of correlations
215 between actual and ASDSpeech estimated ADOS-2 scores. We also performed a similar analysis
216 with NRMSE values, where we assessed whether the actual NRMSE value was smaller than the
217 2.5 percentile of the null distribution. This statistical test, therefore, assessed whether correlation
218 values were higher than expected by chance and NRMSE values were lower than expected by
219 chance.

220 **Data sharing.** The ASDSpeech algorithm source-code and associated dataset are available at
221 <https://github.com/Dinstein-Lab/ASDSpeech>.

222 **Results**

223 Using the data from the 136 ASD children in the training dataset, we examined the relationships
224 between each of the 49 features and ASD symptom severity as defined clinically by the children's
225 ADOS-2 scores. Thirty-one features exhibited significant Pearson correlation coefficients with
226 total ADOS-2 scores (i.e., sum of SA and RRB scores), 31 with ADOS-2 SA scores, and 28 with
227 ADOS-2 RRB scores (Figure 1). While some features, such as the number of vocalizations,
228 exhibited a stronger correlation with SA than RRB score, others, such as mean jitter, exhibited the
229 opposite (Supplementary Figure S4). Hence, different features seem to carry distinct information
230 regarding each of the two core ASD symptoms, demonstrating the potential opportunity for a deep
231 learning algorithm to learn relevant associations.

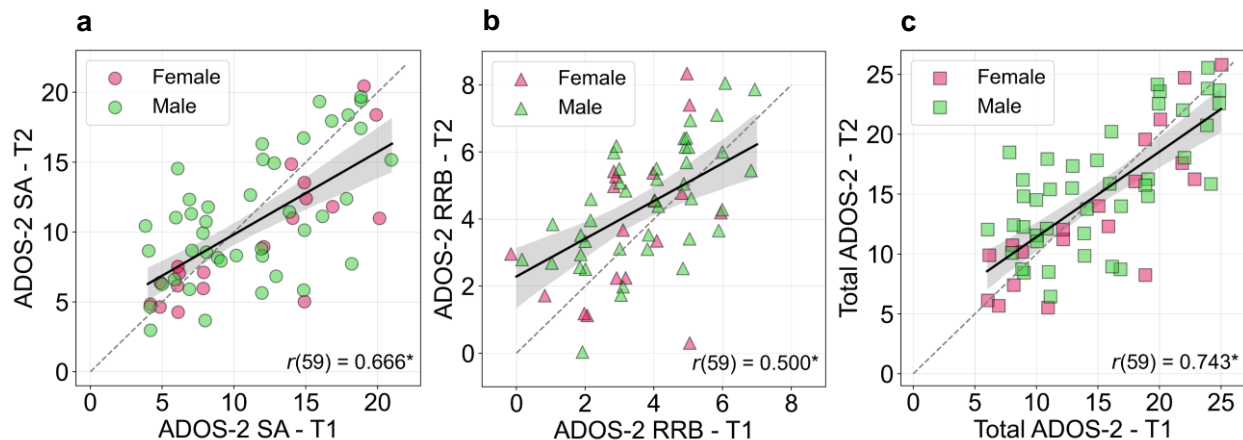


232
 233 **Figure 1. Pearson correlation coefficients between each of the extracted features and ADOS-**
 234 **2 scores from the 136 children in the training dataset.** Correlation coefficients are presented for
 235 total ADOS-2 scores (a), ADOS-2 SA scores (b), and ADOS-2 RRB scores (c). Each color
 236 represents a different group of features. **Asterisks:** significant Pearson correlation (* <math>P < 0.05</math>, ** $P \le$
 237 0.01, *3 $P \le 0.001$, *4 $P \le 0.0001$).

238
 239

240 **Longitudinal stability of ADOS-2 scores**

241 The 61 ASD children in the test dataset exhibited similar ADOS-2 scores across their two
 242 assessments, which were separated by 1-2 years, indicating overall stability in severity over time.
 243 Significant correlations were apparent across first and second assessments for ADOS-2 total ($r(59)$
 244 = 0.743, $P < 0.001$), ADOS-2 SA ($r(59) = 0.666$, $P < 0.001$), and ADOS-2 RRB ($r(59) = 0.5$, $P <$
 245 0.001) scores (Figure 2).



246

247 **Figure 2. Scatter plots demonstrating overall stability in ADOS-2 scores across first and**
248 **second assessments (T1 and T2).** (a) ADOS-2 SA scores. (b) ADOS-2 RRB scores. (c) Total
249 ADOS-2 scores (sum of SA and RRB scores). **Asterisk:** statistical significance of the Pearson
250 correlation coefficient ($P < 0.0001$). **Shaded areas:** 95% confidence intervals. Children located
251 below the diagonal (dashed line) exhibited lower ASD severity at T2 (improvement), while
252 children above the diagonal exhibited the opposite.

253

254 Training and testing the ASDSpeech algorithm

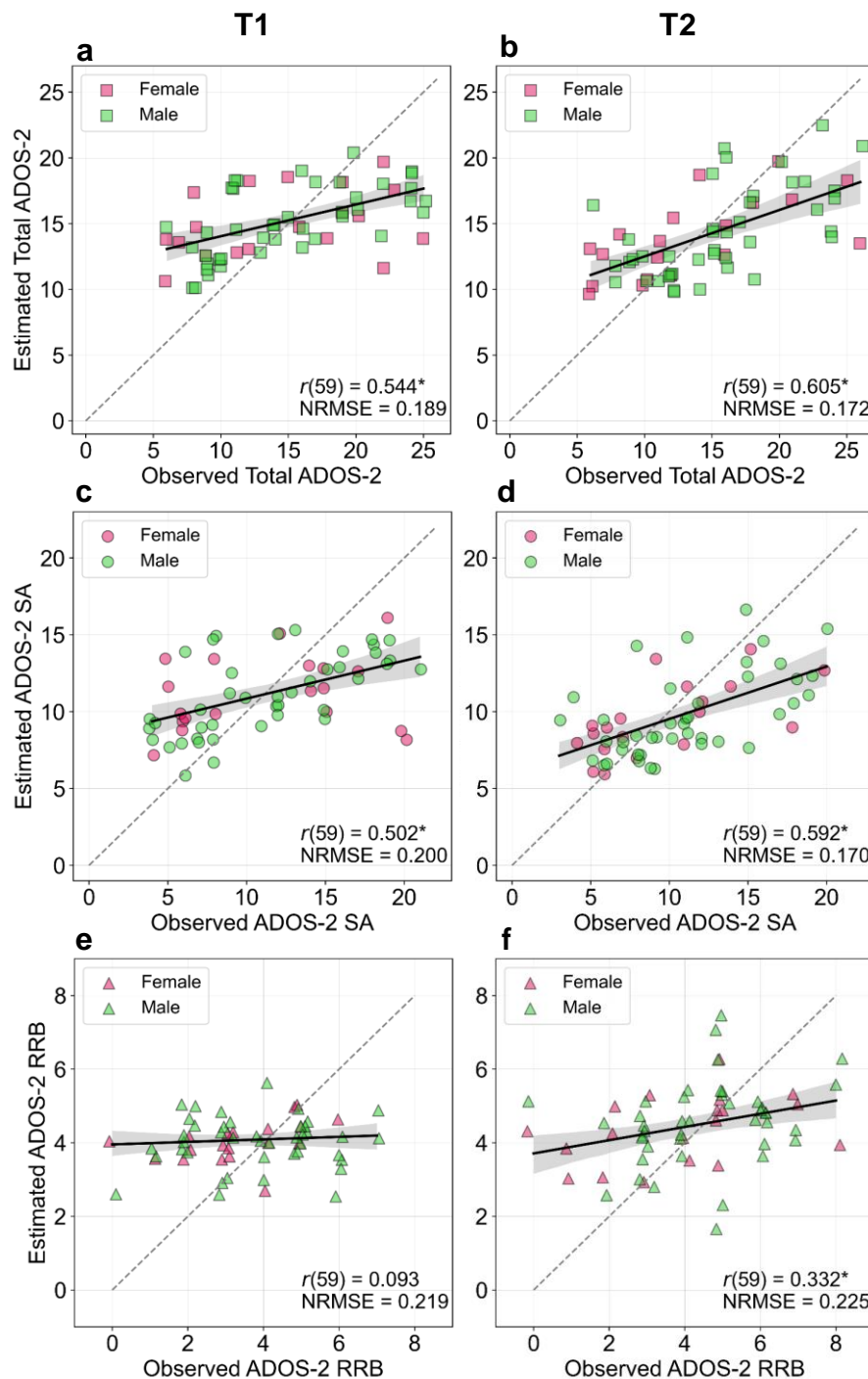
255 We trained the ASDSpeech algorithm with data from 136 ASD children in the training dataset.
256 The algorithm included two separate CNN models that were trained to estimate ADOS-2 SA and
257 RRB scores independently, given that different speech features were associated with each
258 symptom domain. The accuracy of the algorithm was tested with data from two independent
259 ADOS-2 recordings of the 61 children in the testing dataset where ASDSpeech estimated the SA,
260 RRB, and total ADOS-2 (sum of SA and RRB) scores of each child per recording (Figure 3).

261 Estimated total ADOS-2 scores were significantly correlated with actual scores at T1 ($r(59) =$
262 $0.544, P < 0.0001$) and T2 ($r(59) = 0.605, P < 0.0001$). Similarly, estimated ADOS-2 SA scores
263 were significantly correlated with actual scores at T1 ($r(59) = 0.502, P < 0.0001$) and T2 ($r(59) =$
264 $0.592, P < 0.0001$). In contrast, estimated ADOS-2 RRB scores were not significantly correlated
265 with actual RRB scores at T1 ($r(59) = 0.093, P = 0.474$), exhibiting significant correlations only
266 at T2 ($r(59) = 0.332, P = 0.009$) with a relatively weaker effect size.

267 Normalized Root Mean Squared Error (NRMSE) between estimated and actual total ADOS-2
268 scores was significantly smaller than expected by chance when computed at T1 (NRMSE = 0.189,
269 $P < 0.0001$) and T2 (NRMSE = 0.172, $P = 0.0001$). Similarly, NRMSE between estimated and
270 actual ADOS-2 SA scores was significantly smaller than expected by chance when computed at
271 T1 (NRMSE = 0.200, $P < 0.0001$) and T2 (NRMSE = 0.170, $P < 0.0001$). In contrast, NRMSE
272 between estimated and actual ADOS-2 RRB scores was not significantly smaller than expected by

273 chance at T1 (NRMSE = 0.219, $P = 0.460$), exhibiting significant results only at T2 (NRMSE =
274 0.225, $P = 0.006$).

275 The statistical significance of the NRMSE results was determined with a randomization
276 analysis where we randomly shuffled ADOS-2 scores across children before computing NRMSE
277 values. We computed 1,000 random permutations to generate a null NRMSE distribution and
278 assessed statistical significance by determining whether the actual NRMSE value was smaller than
279 the 2.5 percentile of the null distribution (see Methods).



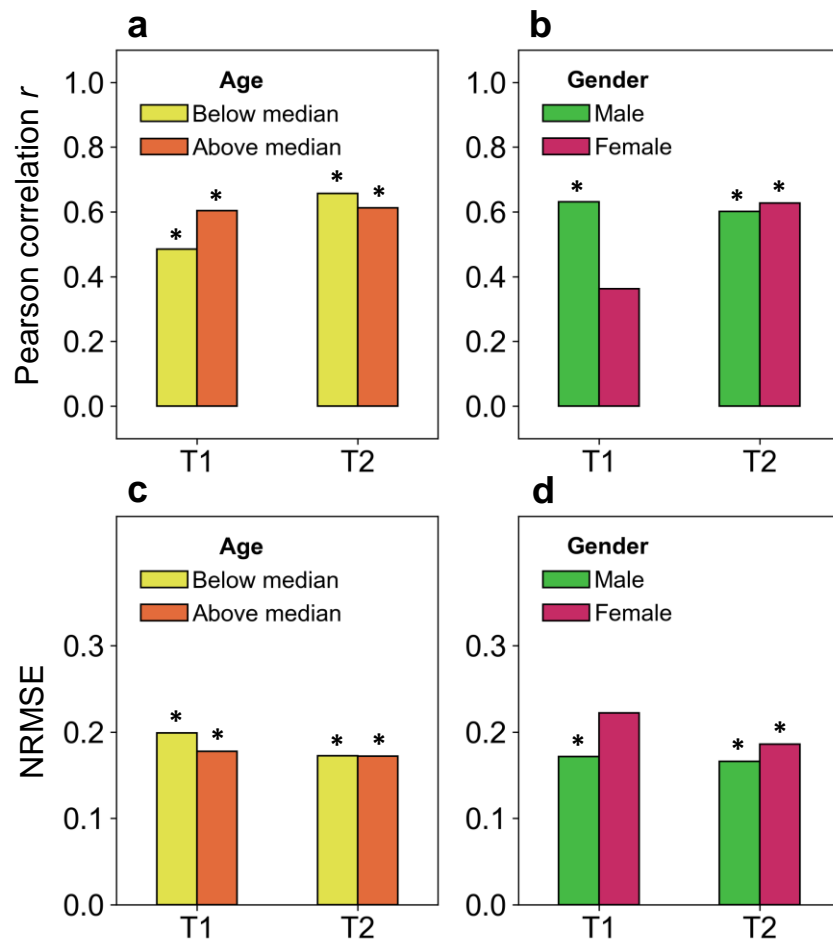
280

281 **Figure 3. Accuracy of ASDSpeech.** Scatter plots demonstrating the fit between estimated and actual scores for the children at T1 (left column) and T2 (right column). (a-b) Total ADOS-2
282 scores (sum of SA and RRB scores). (c-d) ADOS-2 SA scores. (e-f) ADOS-2 RRB scores. Pearson
283 correlation coefficients and NRMSE values are noted in each panel. **Solid line:** Linear fit. **Dashed**
284 **line:** diagonal (unity line). **Asterisks:** statistical significance as determined by randomization test
285 ($P < 0.05$).
286

287 **Differences across age and sex subgroups**

288 Next, we examined whether ASDSpeech accuracy differed across age and sex subgroups
289 (Figure 4). Estimated total ADOS-2 scores were significantly correlated with actual scores when
290 examining children above the median age at T1 ($r(28) = 0.604, P = 0.0004$) or T2 ($r(25) = 0.612,$
291 $P = 0.0007$) and children below the median age at T1 ($r(29) = 0.485, P = 0.006$) or T2 ($r(32) =$
292 $0.657, P < 0.0001$). There were no significant differences in the algorithm's accuracy between
293 younger and older children at T1 ($P = 0.540$) or T2 ($P = 0.780$) as tested with a randomization
294 analysis. Similarly, estimated total ADOS-2 scores were significantly correlated with actual scores
295 when examining males at T1: ($r(40) = 0.631, P < 0.0001$) or T2 ($r(40) = 0.601, P < 0.0001$).
296 Estimated ADOS-2 scores were also significantly correlated with actual scores when examining
297 females at T2 ($r(17) = 0.627, P = 0.004$), but the correlation did not reach statistical significance
298 at T1 ($r(17) = 0.363, P = 0.127$). Nevertheless, there were no significant differences in the
299 algorithm's accuracy between males and females at T1 ($P = 0.198$) or T2 ($P = 0.930$) as tested
300 with a randomization analysis.

301 Comparison of NRMSE across subgroups showed similar results. NRMSE between the
302 estimated and actual ADOS-2 scores was significantly smaller than expected by chance when
303 examining younger children at T1 (NRMSE = 0.199, $P = 0.008$) or T2 (NRMSE = 0.173, $P <$
304 0.0001) as well as older children at T1 (NRMSE = 0.178, $P < 0.0001$) or T2 (NRMSE = 0.172, P
305 < 0.0001). There were no significant differences in the algorithm's accuracy between younger and
306 older children at T1 ($P = 0.434$) or T2 ($P = 0.992$). NRMSE were also significantly smaller than
307 expected by chance when examining males at T1 (NRMSE = 0.172, $P < 0.0001$) or T2 (NRMSE
308 = 0.166, $P < 0.0001$). For females this was the case only at T2 (NRMSE = 0.186, $P = 0.006$) and
309 not at T1 (NRMSE = 0.222, $P = 0.140$). Nevertheless, there were no significant differences in the
310 algorithm's accuracy between males and females at T1 ($P = 0.094$) or T2 ($P = 0.588$) as tested
311 with a randomization test.



312

313 **Figure 4. ASDSpeech accuracy as a function of sex and age at T1 and T2. (a,b)** Pearson
314 correlation values **(c,d)** Normalized Root Mean Squared Error (NRMSE) values. **(a, c)** comparison
315 between younger and older children (median split according to age at each timepoint). **(b, d)**
316 comparison between males and females. **Asterisks:** statistical significance as determined by
317 randomization test ($P < 0.05$).

318 Discussion

319 Our results demonstrate the ability of ASDSpeech to quantify the severity of social symptoms
320 in ASD children from recordings of their speech during ADOS-2 assessments. The algorithm,
321 trained with recordings from 136 ASD children, was able to accurately estimate total ADOS-2 and
322 ADOS-2 SA scores in an entirely independent sample of 61 ASD children, who were recorded at
323 two different developmental timepoints separated by 1-2 years (Figure 3). It is remarkable that
324 ASDSpeech was able to achieve this despite the large heterogeneity in language fluency and
325 speech articulation abilities apparent across ASD children⁴⁵ as well as the large developmental
326 changes that take place in speech abilities during the examined period of early childhood⁴⁶.
327 Moreover, the robust accuracy of ASDSpeech in estimating ADOS-2 SA scores is remarkable

328 given that the social difficulties assessed during the ADOS-2 assessment manifest themselves in
329 behaviors that have little to do with speech including difficulties with eye contact, imitation, joint
330 attention, and other social behaviors^{3,47}. This suggests that combining ASDSpeech with analysis
331 of eye tracking⁴⁸⁻⁵⁰, facial expressions⁵¹, and body movement⁵² data from the same children will
332 enable even higher accuracy and reliability in estimating ASD symptoms.

333 Separate estimation of social and RRB symptoms demonstrated that ASDSpeech was
334 considerably more accurate at estimating social ASD symptoms captured by the ADOS-2 SA
335 scores in contrast to the RRB symptoms captured by the ADOS-2 RRB scores (Figure 3). Note
336 that accurate estimation of total ADOS-2 scores (Figure 3) was likely based on the accurate
337 estimation of SA scores that account for two-thirds of the total scores. We believe there may be
338 several reasons for the more accurate estimation of SA scores. First, the limited range of the
339 ADOS-2 RRB scale (0–8) relative to the SA scale (0–22) may make it difficult for the algorithm
340 to identify differences across children. Indeed, a recent study reported that the limited number of
341 items on the RRB scale resulted in poor scale reliability across participants⁵³. Second, the selected
342 speech features in the current study exhibited weaker correlations with RRB than SA scores
343 (Figure 1). Extraction of additional speech features, such as phrase or intonation repetitions
344 (indicative of echolalia) may improve the accuracy of ADOS-2 RRB score estimates. Regardless,
345 our results motivate separate modeling of social and RRB symptom domains as each of them is
346 likely associated with distinct features of speech.

347 In the current study we estimated raw ADOS-2 scores rather than calibrated severity scores
348 (CSS), which were developed to standardize ASD symptom severity measurements across
349 different ages and language abilities^{30,31}. While ADOS-2 CSS are important for longitudinal
350 assessments of coarse changes in severity^{54,55}, their restricted scoring range (children with ASD
351 receive scores of 4-10) limits the sensitivity of deep learning algorithms in identifying differences
352 across children. By demonstrating that ASDSpeech achieves robust accuracy in estimating raw
353 ADOS-2 SA scores across different age groups and developmental timepoints we show that
354 severity estimations are independent of these factors, thereby justifying the use of raw scores.

355 **Diagnostic classification with speech analysis algorithms**

356 A variety of previous studies have reported that individuals with ASD, on average, speak
357 differently than TD individuals^{4,8-16}. According to these studies, ASD individuals exhibit atypical
358 speech characteristics, including significantly fewer phonemes per utterance¹¹, fewer
359 conversational turns¹³, higher pitch^{9,19}, and larger pitch range and variability^{8,9} than TD children.
360 Differences in these and other speech characteristics have enabled the development of machine
361 and deep learning classification algorithms that can identify ASD and TD individuals with reported
362 accuracy rates of 75-98%¹⁷⁻²³.

363 However, these relatively high classification accuracies are likely to be inflated due to the
364 small sample size of most studies (<40 ASD participants) that are not likely to capture the true
365 heterogeneity of ASD symptoms or speech styles/characteristics of the broad ASD population.
366 Indeed, even “gold standard” clinical tests such as the ADOS-2, exhibit ~80% accuracy in
367 identifying children who will eventually receive an ASD diagnosis⁵⁶. This is because establishing
368 an ASD diagnosis requires clinicians to incorporate additional information from parent interviews
369 and other clinical assessments⁵⁷. Clinicians also report high diagnosis certainty in only ~70% of
370 ASD children because the presentation of ASD symptoms is equivocal in ~30% of cases⁵⁸. These
371 studies suggest an expected upper limit of 70–80% accuracy when attempting to identify ASD
372 using digital phenotyping techniques such as speech analysis. Nevertheless, it is highly
373 encouraging that speech features contain information enabling the separation of ASD and TD
374 children.

375 **Quantifying ASD severity with speech analysis algorithms**

376 A more complex task is to develop machine and deep learning algorithms that can quantify the
377 severity core ASD symptoms. Results presented in the current and previous study from our lab²⁸
378 demonstrated that multiple speech features were significantly correlated with SA and/or RRB
379 ADOS-2 scores (Figure 1), suggesting that distinct combinations of speech features are associated
380 with each of the two core ASD symptoms.

381 Three recent studies have attempted to use these relationships to predict ADOS-2 scores by
382 analyzing speech recordings of ASD individuals^{26–28}. The first trained a synthetic random forest
383 model to estimate ADOS-2 SA scores according to vocalization rate and turn-taking features
384 extracted from ADOS-2 recordings of 33 ASD children. The algorithm was able to predict ADOS-
385 2 SA scores that were significantly correlated with actual scores ($r = 0.634$). The second study
386 utilized a DNN model to estimate four ADOS (first edition) item scores using hundreds of
387 conversational and acoustic features extracted from speech recordings of 88 high-functioning ASD
388 adolescents/adults during an ADOS assessment²⁷. This algorithm was able to estimate scores that
389 exhibited significant Spearman correlations with the actual scores ($\rho = 0.519–0.645$). Finally, in a
390 previous study from our lab²⁸, we demonstrated that a CNN model was able to estimate ADOS-2
391 total scores that were significantly correlated with actual scores ($r = 0.718$) when using 60
392 conversational and acoustic features extracted from speech recordings of 72 children (56 of them
393 with ASD) during ADOS-2 assessment.

394 The current study extends previous work in several critical ways. First, we utilized a
395 considerably larger dataset (258 ADOS-2 recordings) that was at least three times larger than the
396 ones used to date. This was important for training ASDSpeech with speech recordings from a large
397 cohort with heterogeneous language abilities. Second, the 61 ASD children in our testing dataset
398 were recorded twice during two ADOS-2 assessments separated by 1–2 years. This enabled us to
399 test the robustness of ASDSpeech across two developmental timepoints. Third, we trained

400 ASDSpeech to estimate ADOS-2 SA and ADOS-2 RRB scores using separate CNN models. The
401 results demonstrated that this separation was critical with accurate performance apparent primarily
402 for the ADOS-2 SA scores. Fourth, the large sample size enabled us to demonstrate that
403 ASDSpeech accuracy was similar across age and sex subgroups. Fifth, the recordings utilized in
404 the current study were performed over a 6-year period in a busy public healthcare medical center
405 that services a population of ~1 million people. Recordings were performed with a wall mounted
406 microphone (see Methods) in “real world” noisy conditions (e.g., announcement system in the
407 hallway). This demonstrates the robustness of ASDSpeech to variable recording conditions.

408 ASDSpeech achieved similar accuracy to that reported in previous studies. The important
409 advance in the current study is in demonstrating that this accuracy is robust to age and
410 developmental stage of the examined children when examining a large heterogeneous population
411 within an active clinical setting. Most importantly, we openly share ASDSpeech and its associated
412 dataset with the research community.

413 **Limitations**

414 The current study had several limitations. First, we did not examine the language content of the
415 recordings, which is likely to improve the estimation of ASD symptom severity^{4,24}. Second, we
416 did not identify echolalia, crying, or shouting events that are likely to be informative of RRB
417 symptoms. Indeed, our weaker results estimating RRB scores suggest that different speech features
418 are necessary for estimating severity in this domain. Third, we did not apply any noise reduction
419 or multi-speaker analysis techniques to improve the quality of the analyzed vocal segments.
420 Finally, our sample had a 4:1 male to female ratio, which is equivalent to the sex ratio in the
421 national ASD population of Israel⁵⁹. Hence, higher ASDSpeech accuracy for males at T1 may be
422 due to the larger number of males in the training and testing datasets. This could be rectified by
423 future studies.

424 **Conclusions**

425 This study adds to accumulating evidence demonstrating that speech recordings contain reliable
426 information about the social symptom severity of ASD children. We demonstrate the ability of the
427 ASDSpeech algorithm to quantify these symptoms in a robust manner across two developmental
428 timepoints with recordings that were performed within a busy community healthcare center. We
429 openly share the algorithm and its associated dataset for further use, testing, and development by
430 the research community and are confident that future versions of the algorithm will achieve even
431 higher and more robust accuracy rates, yielding a transformative new tool for clinical and basic
432 ASD research.

433 **References**

- 434 1 Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. *The*
435 *Lancet* 2018; **392**: 508–520.
- 436 2 Gabbay-Dizdar N, Ilan M, Meiri G, Faroy M, Michaelovski A, Flusser H *et al.* Early
437 diagnosis of autism in the community is associated with marked improvement in social
438 symptoms within 1–2 years. *Autism* 2022; **26**: 1353–1363.
- 439 3 Tager-Flusberg H, Kasari C. Minimally Verbal School-Aged Children with Autism
440 Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research* 2013; **6**: 468–
441 478.
- 442 4 Salem AC, MacFarlane H, Adams JR, Lawley GO, Dolata JK, Bedrick S *et al.* Evaluating
443 atypical language in autism using automated language measures. *Sci Rep* 2021; **11**.
444 doi:10.1038/S41598-021-90304-5.
- 445 5 Chi NA, Washington P, Kline A, Husic A, Hou C, He C *et al.* Classifying Autism From
446 Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison
447 Study. *JMIR Pediatr Parent* 2022; **5**: e35406.
- 448 6 Mostek J. Cognitive Development and Language Acquisition in Autistic Children. *Science*
449 *Insights* 2022; **41**: 719–724.
- 450 7 Loukusa S. Autism Spectrum Disorder. In: Cummings L (ed). *Handbook of Pragmatic*
451 *Language Disorders*. Springer International Publishing: Cham, 2021, pp 45–78.
- 452 8 Bonnef YS, Levanon Y, Dean-Pardo O, Lossos L, Adini Y. Abnormal Speech Spectrum
453 and Increased Pitch Variability in Young Autistic Children. *Front Hum Neurosci* 2011; **4**:
454 237.
- 455 9 Asghari SZ, Farashi S, Bashirian S, Jenabi E. Distinctive prosodic features of people with
456 autism spectrum disorder: a systematic review and meta-analysis study. *Sci Rep* 2021; **11**:
457 23093.
- 458 10 Oller DK, Niyogi P, Gray S, Richards JA, Gilkerson J, Xu D *et al.* Automated vocal analysis
459 of naturalistic recordings from children with autism, language delay, and typical
460 development. *Proceedings of the National Academy of Sciences* 2010; **107**: 13354–13359.
- 461 11 Moffitt JM, Ahn YA, Custode S, Tao Y, Mathew E, Parlade M *et al.* Objective measurement
462 of vocalizations in the assessment of autism spectrum disorder symptoms in preschool age
463 children. *Autism Research* 2022; : 1–10.
- 464 12 Ferguson EF, Nahmias AS, Crabbe S, Liu T, Mandell DS, Parish-Morris J. Social language
465 opportunities for preschoolers with autism: Insights from audio recordings in urban
466 classrooms. *Autism* 2020; **24**: 1232–1245.

- 467 13 Warren SF, Gilkerson J, Richards JA, Oller DK, Xu D, Yapanel U *et al.* What Automated
468 Vocal Analysis Reveals About the Vocal Production and Language Learning Environment
469 of Young Children with Autism. *J Autism Dev Disord* 2010; **40**: 555–569.
- 470 14 Warlaumont AS, Richards JA, Gilkerson J, Oller DK. A Social Feedback Loop for Speech
471 Development and Its Reduction in Autism. *Psychol Sci* 2014; **25**: 1314–1324.
- 472 15 Plumb AM, Wetherby AM. Vocalization Development in Toddlers With Autism Spectrum
473 Disorder. *Journal of Speech, Language, and Hearing Research* 2013; **56**: 721–734.
- 474 16 Tenenbaum EJ, Carpenter KL, Sabatos-DeVito M, Hashemi J, Vermeer S, Sapiro G *et al.*
475 A Six-Minute Measure of Vocalizations in Toddlers with Autism Spectrum Disorder.
476 *Autism Research* 2020; **13**: 1373–1382.
- 477 17 Pokorny FB, Schuller BW, Marschik PB, Brueckner R, Nyström P, Cummins N *et al.*
478 Earlier identification of children with autism spectrum disorder: An automatic vocalisation-
479 based approach. *Proceedings of the Annual Conference of the International Speech*
480 *Communication Association, INTERSPEECH* 2017; **2017-Augus**: 309–313.
- 481 18 Mohanta A, Mittal VK. Classifying Speech of ASD Affected and Normal Children Using
482 Acoustic Features. In: *2020 National Conference on Communications (NCC)*. IEEE, 2020,
483 pp 1–6.
- 484 19 Mohanta A, Mittal VK. Analysis and classification of speech sounds of children with autism
485 spectrum disorder using acoustic features. *Comput Speech Lang* 2022; **72**: 101287.
- 486 20 Asgari M, Shafran I. Robust and Accurate Features for Detecting and Diagnosing Autism
487 Spectrum Disorders. *Proceedings of the Annual Conference of the International Speech*
488 *Communication Association, INTERSPEECH* 2013; : 191–194.
- 489 21 Yankowitz LD, Schultz RT, Parish-Morris J. Pre- and Paralinguistic Vocal Production in
490 ASD: Birth Through School Age. *Curr Psychiatry Rep* 2019; **21**: 126.
- 491 22 Lee S, Yeo EJ, Kim S, Chung M. Knowledge-driven speech features for detection of
492 Korean-speaking children with autism spectrum disorder*. *Phonetics and Speech Sciences*
493 2023; **15**: 53–59.
- 494 23 Briend F, David C, Silleresi S, Malvy J, Ferré S, Latinus M. Voice acoustics allow
495 classifying autism spectrum disorder with high accuracy. *Transl Psychiatry* 2023; **13**: 250.
- 496 24 MacFarlane H, Salem AC, Chen L, Asgari M, Fombonne E. Combining voice and language
497 features improves automated autism detection. *Autism Res* 2022; **15**: 1288–1300.
- 498 25 Lord C, Rutter M, Di Lavore P, Risi S, Gotham K, Bishop S. Autism and Diagnostic
499 Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1-4. 2012.

- 500 26 Sadiq S, Castellanos M, Moffitt J, Shyu M, Perry L, Messinger D. Deep Learning Based
501 Multimedia Data Mining for Autism Spectrum Disorder (ASD) Diagnosis. In: *2019*
502 *International Conference on Data Mining Workshops (ICDMW)*. 2019, pp 847–854.
- 503 27 Chen C-P, Gau SS-F, Lee C-C. Learning Converse-Level Multimodal Embedding to Assess
504 Social Deficit Severity for Autism Spectrum Disorder. In: *2020 IEEE International*
505 *Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp 1–6.
- 506 28 Eni M, Dinstein I, Ilan M, Menashe I, Meiri G, Zigel Y. Estimating Autism Severity in
507 Young Children From Speech Signals Using a Deep Neural Network. *IEEE Access* 2020;
508 **8**: 139489–139500.
- 509 29 Asghari SZ, Farashi S, Bashirian S, Jenabi E. Distinctive prosodic features of people with
510 autism spectrum disorder: a systematic review and meta-analysis study. *Sci Rep* 2021; **11**:
511 23093.
- 512 30 Esler AN, Bal VH, Guthrie W, Wetherby A, Weismer SE, Lord C. The Autism Diagnostic
513 Observation Schedule, Toddler Module: Standardized Severity Scores. *J Autism Dev Disord*
514 2015; **45**: 2704–2720.
- 515 31 Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in
516 autism spectrum disorders. *J Autism Dev Disord* 2009; **39**: 693–705.
- 517 32 Dinstein I, Arazi A, Golan HM, Koller J, Elliott E, Gozes I *et al*. The National Autism
518 Database of Israel: a Resource for Studying Autism Risk Factors, Biomarkers, Outcome
519 Measures, and Treatment Efficacy. *Journal of Molecular Neuroscience* 2020; **70**: 1303–
520 1312.
- 521 33 Meiri G, Dinstein I, Michaelowski A, Flusser H, Ilan M, Faroy M *et al*. The Negev Hospital-
522 University-Based (HUB) Autism Database. *J Autism Dev Disord* 2017; **47**: 2918–2926.
- 523 34 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*.
524 5th ed. American Psychiatric Publishing: Arlington, VA, 2013
525 doi:10.1176/appi.books.9780890425596.744053.
- 526 35 Lord C, Rutter M, Di Lavore P, Risi S, Gotham K, Bishop S. Autism and Diagnostic
527 Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1-4. 2012.
- 528 36 Padmini P, Gupta D, Zakariah M, Alotaibi YA, Bhowmick K. A Simple Speech Production
529 System Based on Formant Estimation of a Tongue Articulatory System Using Human
530 Tongue Orientation. *IEEE Access* 2021; **9**: 4688–4710.
- 531 37 Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J *et al*. Characteristics
532 of motor speech phenotypes in multiple sclerosis. *Mult Scler Relat Disord* 2018; **19**: 62–69.

- 533 38 Boersma P, van Heuven V. Speak and unSpeak with Praat. *Glott International* 2001; **5**: 341–
534 347.
- 535 39 G Pillai L, Sherly E. A Deep Learning Based Evaluation of Articulation Disorder and
536 Learning Assistive System for Autistic Children. *International Journal on Natural*
537 *Language Computing* 2017; **6**: 19–36.
- 538 40 Tamarit L, Goudbeek M, Scherer K. Spectral Slope Measurements in Emotionally
539 Expressive Speech. In: *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on*
540 *Speech Analysis and Processing for Knowledge Discovery*. 2008, pp 1–4, paper 007.
- 541 41 Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C *et al*. The Geneva
542 Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective
543 Computing. *IEEE Trans Affect Comput* 2016; **7**: 190–202.
- 544 42 Taqi AM, Awad A, Al-Azzo F, Milanova M. The Impact of Multi-Optimizers and Data
545 Augmentation on TensorFlow Convolutional Neural Network Performance. *Proceedings -*
546 *IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018*
547 2018; **April**: 140–145.
- 548 43 Prechelt L. Early Stopping — But When? In: *Neural Networks: Tricks of the Trade*. 2012,
549 pp 53–67.
- 550 44 MacFarlane H, Salem AC, Bedrick S, Dolata JK, Wiedrick J, Lawley GO *et al*. Consistency
551 and reliability of automated language measures across expressive language samples in
552 autism. *Autism Research* 2023; **16**: 802–816.
- 553 45 Tager-Flusberg H. Defining language phenotypes in autism. *Clin Neurosci Res* 2006; **6**:
554 219–224.
- 555 46 Tager-Flusberg H, Rogers S, Cooper J, Landa R, Lord C, Paul R *et al*. Defining Spoken
556 Language Benchmarks and Selecting Measures of Expressive Language Development for
557 Young Children With Autism Spectrum Disorders. *Journal of Speech, Language, and*
558 *Hearing Research* 2009; **52**: 643–652.
- 559 47 Vivanti G, Bottema-Beutel K, Turner-Brown L. Understanding and Addressing Restricted
560 and Repetitive Behaviors in Children with Autism. In: *Clinical Guide to Early Interventions*
561 *for Children with Autism*. 2020, pp 61–77.
- 562 48 Avni I, Meiri G, Bar-Sinai A, Reboh D, Manelis L, Flusser H *et al*. Children with autism
563 observe social interactions in an idiosyncratic manner. *Autism Research* 2020; **13**: 935–946.
- 564 49 Chong E, Clark-Whitney E, Southerland A, Stubbs E, Miller C, Ajodan EL *et al*. Detection
565 of eye contact with deep neural networks is as accurate as human experts. *Nat Commun*
566 2020; **11**: 6386.

- 567 50 Jones W, Klaiman C, Richardson S, Aoki C, Smith C, Minjarez M *et al.* Eye-Tracking-
568 Based Measurement of Social Visual Engagement Compared With Expert Clinical
569 Diagnosis of Autism. *JAMA* 2023; **330**: 854–865.
- 570 51 Perochon S, Di Martino JM, Carpenter KLH, Compton S, Davis N, Eichner B *et al.* Early
571 detection of autism using digital behavioral phenotyping. *Nat Med* 2023; **29**: 2489–2497.
- 572 52 Budman I, Meiri G, Ilan M, Faroy M, Langer A, Reboh D *et al.* Quantifying the social
573 symptoms of autism using motion capture. *Sci Rep* 2019; **9**: 7712.
- 574 53 Frazier TW, Whitehouse AJO, Leekam SR, Carrington SJ, Alvares GA, Evans DW *et al.*
575 Reliability of the Commonly Used and Newly-Developed Autism Measures. *J Autism Dev*
576 *Disord* 2023. doi:10.1007/s10803-023-05967-y.
- 577 54 Gabbay-Dizdar N, Ilan M, Meiri G, Faroy M, Michaelovski A, Flusser H *et al.* Early
578 diagnosis of autism in the community is associated with marked improvement in social
579 symptoms within 1–2 years. *Autism* 2022; **26**: 1353–1363.
- 580 55 Waizbard-Bartov E, Ferrer E, Young GS, Heath B, Rogers S, Wu Nordahl C *et al.*
581 Trajectories of Autism Symptom Severity Change During Early Childhood. *J Autism Dev*
582 *Disord* 2021; **51**: 227–242.
- 583 56 Maddox BB, Brodtkin ES, Calkins ME, Shea K, Mullan K, Hostager J *et al.* The Accuracy
584 of the ADOS-2 in Identifying Autism among Adults with Complex Psychiatric Conditions.
585 *J Autism Dev Disord* 2017; **47**: 2703.
- 586 57 Bishop SL, Lord C. Commentary: Best practices and processes for assessment of autism
587 spectrum disorder - the intended role of standardized diagnostic instruments. *J Child*
588 *Psychol Psychiatry* 2023; **64**. doi:10.1111/JCPP.13802.
- 589 58 Klaiman C, White S, Richardson S, McQueen E, Walum H, Aoki C *et al.* Expert Clinician
590 Certainty in Diagnosing Autism Spectrum Disorder in 16-30-Month-Olds: A Multi-site
591 Trial Secondary Analysis. *J Autism Dev Disord* 2022. doi:10.1007/S10803-022-05812-8.
- 592 59 Raz R, Weisskopf MG, Davidovitch M, Pinto O, Levine H. Differences in autism spectrum
593 disorders incidence by sub-populations in Israel 1992-2009: a total population study. *J*
594 *Autism Dev Disord* 2015; **45**: 1062–1069.

595

596 **Acknowledgments**

597 This study was supported by the Israeli Science Foundation (Grant no. 1150/20) and the Israel
598 Ministry of Science & Technology (Grant no. 3-17422).

599 **Author contributions**

600 M.E. collected the data, performed the experiments, built the models, analyzed the data, and wrote
601 the paper. I.D., and Y.Z. designed the study, guided data collection and analysis, and wrote the
602 paper. M.I., A.M., H.M.G., G.M., and I.M. contributed to participant recruitment, behavioral
603 assessments, data collection, and interpretation of the findings. All authors approved the final
604 manuscript.

605 **Additional information**

606 **Competing interests.** The authors declare no competing interests.