Quantifying longitudinal changes in core autism symptoms using automated speech analysis

Marina Eni, Michal Ilan, Analya Michaelovski, Hava Golan, Gal Meiri, Idan Menashe, Ilan Dinstein, and Yaniv Zigel

Abstract

Social communication problems, which are a core symptom of autism, may be evident in the speech characteristics of autistic individuals. Here, we examined acoustic and conversational features extracted from audio recordings of Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) assessments of 1.5-7-year-old children. We trained a deep neural network algorithm to estimate autism severity (i.e., ADOS-2 scores) from recordings of 146 children and tested its accuracy with independent recordings from 62 additional children who completed two ADOS-2 assessments, separated by 1-2 years. Estimated ADOS-2 social affect scores in the test set were significantly correlated with true scores at each time-point ($r(62)=0.442\text{-}0.575$, $P<0.001$), and estimated changes across time-points were significantly correlated with true changes ($r(62)=0.343$, $P=0.011$). The presented algorithm learned to estimate social symptom severity from speech recordings of one autism group and was able to accurately estimate severity changes in an entirely independent group. While accuracy needs to be further improved by training with larger datasets, these results demonstrate the remarkable utility of speech analysis algorithms in estimating autism risk and severity changes over time.

Introduction

Autism Spectrum Disorder (ASD) is diagnosed by the presence of social communication difficulties and the existence of Restricted and Repetitive Behaviors (RRBs) [1]. Most ASD children exhibit language delays during early childhood [2], with 25-30% remaining minimally verbal (i.e., use <50 words) throughout childhood [3]. However, core ASD symptoms are not necessarily evident in the amount of speech produced by an individual and may instead be evident in the way they speak. Some ASD children exhibit poorer fluency [4], echolalia (i.e., speech repetition) [5], mix pronouns [6], and use atypical articulation and prosody [7], [8] that are apparent in the acoustic features of their vocalizations [9], [10]. Previous studies have reported, for example, that verbal ASD children tend to speak with higher pitch and larger pitch variability than Typically Developing (TD) children [8], [9].

Recent studies have utilized speech analysis techniques to identify speech characteristics that differ in ASD children. These studies have reported that ASD children exhibit significantly fewer phoneme vocalizations [11], fewer vocalizations per minute [12], [13], fewer conversational turns (i.e., reciprocating in a conversation) [11], [13], [14], more non-speech vocalizations [12], [14], more distressed vocalizations (crying, screaming) [15], and a lower ratio of syllables to vocalizations [16] than TD children. Additional studies have trained machine learning and deep learning algorithms to classify ASD and TD children based on extracted speech features [17]–[24] and have reported classification accuracies of 75% to 98% [19]. Different studies used different speech characteristics, including linguistic features such as vocabulary and fluency [24] and

acoustic features such as pitch [18]–[20], [22], [23], jitter [20], [23], shimmer [20], [23], energy [18], [19], Zero-Crossing Rate (ZCR) [18], [19], Mel-Frequency Cepstral Coefficients (MFCCs) [19], Linear Predictive Coefficients (LPCs) [19], formants [18], [19], [23], speech rate [22], and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (88 acoustic features commonly used in speech analyses) [17], [21], [22]. Some of these studies utilized classical machine learning techniques, such as Support Vector Machine (SVM) classification [18], [20], [22], [24] or K-Nearest Neighbors (KNN) and K-Means clustering [18], [23] while others used Probabilistic Neural Network (PNN) [19] and Bidirectional Long Short-Term Memory (BLSTM) [17] deep learning algorithms.

Note that the relatively high classification accuracies reported above are likely to be inflated due to small samples (mostly <40 participants in the ASD group) that are not likely to capture the true heterogeneity of ASD symptoms or speech styles/characteristics in the broad ASD population. Indeed, even "gold standard" clinical tests such as the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2), a semi-structured assessment where clinicians score the behavior of children during specific tasks/games, exhibit ~80% accuracy in identifying children who will eventually receive an ASD diagnosis [25]. This is because establishing the ASD diagnosis requires clinicians to incorporate additional information that is gathered in parent interviews and other clinical assessments/tests in addition to the ADOS-2 [26]. Clinicians also report high diagnosis certainty in only ~70% of ASD cases at early ages because the presentation of ASD symptoms is equivocal in ~30% of cases [27]. These studies suggest an expected upper limit of 70-80% accuracy when attempting to identify ASD using digital phenotyping techniques such as speech analysis. Nevertheless, it is highly encouraging that speech features contain information enabling the separation of ASD and TD children.

Further extension of this research has focused on estimating ASD severity from extracted speech features. Studies have demonstrated that multiple speech features are significantly correlated with ADOS-2 scores [11], [28] and three recent studies have suggested that it may be possible to use machine and deep learning algorithms to estimate ASD severity from speech recordings of ASD children [28]–[30]. The first extracted vocalization rates and durations from ADOS-2 recordings of 33 ASD children and trained a synthetic random forest model to estimate their ADOS-2 Social Affect (SA) scores, which measure social communication difficulties. The model predicted ADOS-2 SA scores that were significantly correlated with the actual scores (r = 0.634) when tested on the same recordings used for training (i.e., no independent testing set) [29]. The second study [30] extracted hundreds of conversational, acoustic, and lexical speech features from ADOS recordings of 88 high-functioning adolescents/adults with ASD. A Deep Neural Network (DNN) was used to estimate scores of four specific ADOS items that quantify the ability to maintain a mature social conversation. These items' estimated and actual scores exhibited significant Spearman correlations, on average, in a 5-fold cross-validation training-testing scheme (ρ = 0.519-0.645). Finally, in the third study, from our lab, we extracted acoustic features such as pitch and energy, and conversational features such as turn-taking and speech rate from 72 ADOS-2 recordings of 2-6-year-old children, 56 of them with ASD [28]. We trained a Convolutional Neural Network (CNN) model on data from 51 children and tested on the remaining 21 children, performing this analysis 50 times with random selections of train-test sub-samples. The trained models were able to predict ADOS-2 scores that were, on average, significantly correlated with actual ADOS-2 scores (mean r = 0.718 ± 0.093).

While these results are encouraging, studies to date have used relatively small samples that are often biased to a specific sub-group of ASD participants (e.g., high-functioning adolescents/adults). In addition, previous studies did not compare the ability of deep learning models to successfully estimate the severity of social ASD symptoms versus RRB symptoms. While one may expect both social difficulties and RRB behaviors to impact how an individual speaks, the two symptom domains are likely to impact speech in different ways that may be associated with distinct speech features and characteristics. Indeed, in our previous study, we demonstrated that different speech features were correlated with ADOS-2 SA versus RRB scores [28].

Ultimately, ASD speech analysis algorithms aspire to yield robust ASD severity measures across multiple time-points and identify changes over time or following intervention [31]. Hence, our study had three main goals. First, to determine the accuracy of our speech analysis algorithm in robustly estimating ADOS-2 scores of ASD children recorded in each of two assessments separated by approximately one year. Second, to determine the algorithm's accuracy in estimating changes in ADOS-2 scores over time. Third, to compare the algorithm's accuracy in estimating ADOS-2 SA versus RRB scores (i.e., social difficulties versus RRB symptoms). To the best of our knowledge, this is the first study to examine sensitivity to changes in ASD symptom severity that are known to happen over longer periods of time (>one year) [2], [32]. To perform this study, we created the largest speech recording dataset available to date, which contained 102,987 vocalizations from 208 ASD children recorded in 270 ADOS-2 assessments. Speaker diarization in this relatively large dataset was performed manually to ensure accuracy. We expected this dataset to capture more of the large heterogeneity in symptoms and behaviors that are common in ASD children [33].

Results

A total of 102,987 child vocalizations (including speech, cries, and screams) with a net duration of 18.63 hours were manually identified in the audio recordings of 208 ASD children and analyzed in the current study. Of these children, 146 participated in a single ADOS-2 assessment, and 62 participated in two ADOS-2 assessments separated by 1-2 years. We extracted 49 features of speech from each ADOS-2 recording. These included acoustic features (e.g., pitch, jitter, formants, bandwidth, energy, voicing, and spectral slope) and conversational features (e.g., mean vocalization duration and total number of vocalizations). We then assessed the utility of these features for estimating ADOS-2 scores and their change over time.

Association between acoustic features and autism severity

We computed Pearson correlation coefficients between each of the extracted speech features and the ADOS-2 scores of the 146 ASD children who participated in a single ADOS-2 assessment and made up the training set for the models below. Of all 49 features, 25 features exhibited significant correlations with total ADOS-2 scores (i.e., sum of SA and RRB scores), 30 with ADOS-2 SA scores, and 21 with ADOS-2 RRB scores (Fig. 1). Most of the features that exhibited significant correlations with total ADOS-2 scores in our previous study [28], which was performed with a different smaller sample, were also correlated in the current study (Supplementary Table S1). This suggests that some speech features indeed carry reliable information regarding symptom severity. While some features, such as the number of vocalizations, exhibited a stronger correlation with

SA than RRB score, others, such as mean jitter, exhibited the opposite (Supplementary Fig. S1). Hence, different features seem to carry distinct information regarding each of the two core ASD symptom domains.

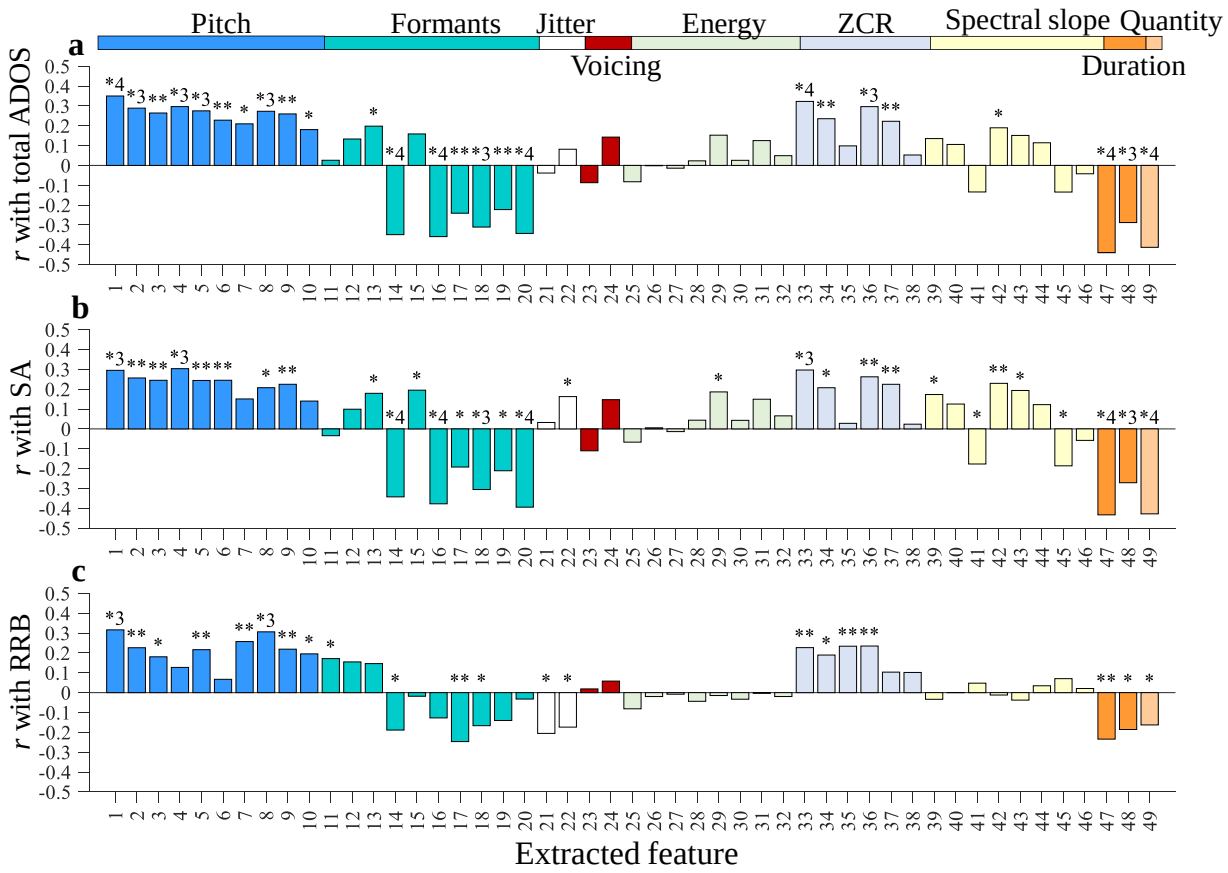

Fig. 1. Pearson correlation coefficients between each of the extracted features and ADOS-2 scores in the 146 children who completed a single ADOS-2 assessment. Correlation coefficients are presented for total ADOS-2 scores (**a**), ADOS-2 SA scores (**b**), and ADOS-2 RRB scores (**c**). Each color represents a different group of features. **Asterisks**: statistical significance of the Pearson correlation coefficient (∗*P* value < 0.05, ∗∗ *P* value ≤ 0.01, ∗3 *P* value ≤ 0.001, ∗4 *P* value ≤ 0.0001).

Changes over time

Next, we examined how THE scores of the 62 children who completed two ADOS-2 assessments changed over time. Total, SA, and RRB scores were significantly correlated across time-points ($r(62)$= 0.5-0.71, $P < 0.001$), indicating that children tended to exhibit similar scores across time-points (Fig. 2). Nevertheless, some children demonstrated changes in the severity of symptoms for better (lower scores at T2) or worse (the opposite). Note that changes in the social domain (ADOS-2 SA) were not correlated with changes in the RRB domain ($r(62)$ = -0.093, P = 0.471, Fig. 2**d**), demonstrating the independence of developmental changes in the two symptom domains.

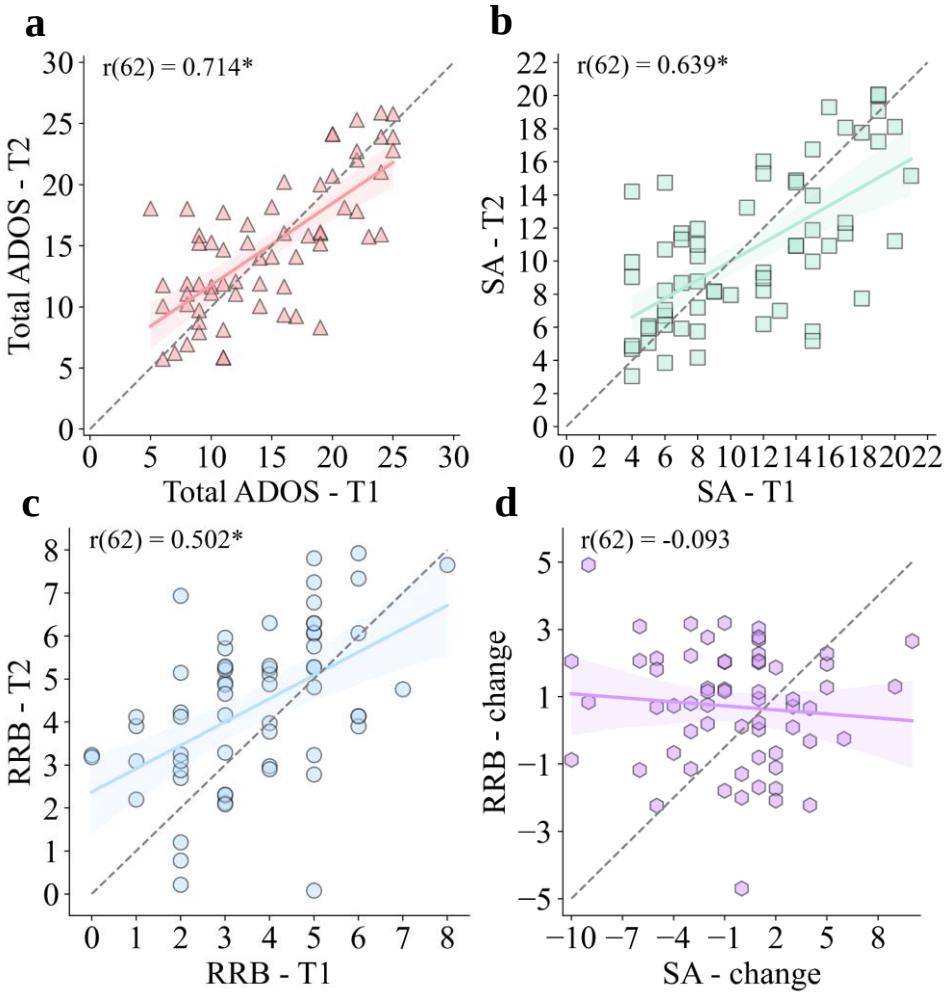

Fig. 2. Scatter plots demonstrating changes in ADOS-2 scores across time-points (T1 and T2). (**a**) Total ADOS-2 scores. (**b**) ADOS-2 SA scores. (**c**) ADOS-2 RRB scores. (**d**) Change in RRB (T2-T1) vs change in SA. Each circle represents a single child from the test set of 62 children. **Asterisk**: statistical significance of the Pearson correlation coefficient (*P* value < 0.001**). Shaded areas**: present 95% confidence intervals. Children located below the diagonal (dashed line) exhibited lower ASD severity at T2 (improvement), while children above the diagonal exhibited the opposite.

Autism Severity Estimation (ASE)

We trained two different ASE algorithms with data from the 146 ASD children who completed a single ADOS-2 assessment. The first algorithm included two separate CNN models, which learned to estimate SA and RRB scores, respectively. The second algorithm included a single CNN model, which learned to estimate total ADOS-2 scores. Once training was complete, the accuracy of each algorithm was tested with data from independent recordings of the 62 children who completed two ADOS assessments. We tested the accuracy of each algorithm with the data from the first and second assessments separately and then tested the accuracy of estimating longitudinal changes (i.e., T2-T1 scores).

To assess the accuracy of the models in a robust manner, we trained and tested each model repeatedly over 500 iterations, where we randomly selected different speech segments from each audio recording when extracting the feature matrices (see methods). We then computed the Pearson correlation coefficient between the estimated ADOS-2 scores and the actual scores in each iteration (see example in Supplementary Fig. S2) and reported their mean and standard deviation across 500 iterations (Fig. 3).

This revealed significant correlations, on average, between the estimated and actual ADOS-2 SA scores at T1 ($r(62) = 0.442 \pm 0.059$, $P<0.001$) and T2 ($r(62) = 0.575 \pm 0.047$, $P < 0.001$) as well as the estimated and actual SA score changes, T2-T1($r(62) = 0.343 \pm 0.078$, $P = 0.011$). In contrast, the estimated ADOS-2 RRB scores were not significantly correlated with actual RRB scores at T1 ($r(62) = -0.035 \pm 0.118$, $P = 0.580$) or T2 ($r(62) = 0.227 \pm 0.123$, $P = 0.102$), and correlations between the estimated and predicted changes in RRB scores were also not significant ($r(62) = -0.034 \pm 0.078$, $P = 0.595$). We also examined Spearman's rank correlation coefficients and found equivalent results (see Supplementary Table S2).

Similarly to the ADOS-2 SA scores, total ADOS-2 estimated scores were significantly correlated with true scores at T1 when summing SA and RRB scores that were estimated separately ($r(62) = 0.410 \pm 0.076$, $P = 0.002$) or when estimating the total ADOS-2 ($r(62) = 0.401 \pm 0.076$, $P = 0.003$). Significant correlations were also apparent at T2 for separately estimated ($r(62) = 0.574 \pm 0.051$, $P < 0.0001$) or total estimated ($r(62) = 0.576 \pm 0.049$, $P < 0.0001$) scores. Finally, significant correlations were apparent for changes over time when estimating SA and RRB scores separately ($r(62) = 0.253 \pm 0.081$, $P = 0.048$) or when estimating the total ADOS-2 scores ($r(62) = 0.260 \pm 0.078$, $P = 0.043$). Note that the statistical significance of all correlations was determined using permutation tests (see methods for details).

Differences across age and sex subgroups

Significant differences in ASE accuracy were apparent across age and sex subgroups of children when using the total ADOS-2 ASE algorithm (Fig. 4). Pearson correlation coefficients were significantly larger for boys in comparison to girls at T1 ($t(998) = 13.91$, $P < 0.0001$) and T2 ($t(998) = 18.65$, $P < 0.0001$). Estimated ADOS-2 total scores were significantly correlated with true scores for boys at T1 ($r(43) = 0.43 \pm 0.08$, $P = 0.006$) and T2 ($r(43) = 0.62 \pm 0.05$, $P < 0.0001$). They were also significant for girls at T2 ($r(19) = 0.52 \pm 0.11$, $P = 0.026$) but not significant at T1 ($r(19) = 0.34 \pm 0.12$, $P = 0.113$).

Correlations were also significantly larger for children younger than the median age compared to older children at T1 ($t(998) = 4.55$, $P < 0.0001$) and T2 ($t(998) = 29.43$, $P < 0.0001$). Estimated ADOS-2 total scores were significantly correlated with true scores at T1 for relatively younger ($r(31) = 0.42 \pm 0.12$, $P = 0.032$) and older ($r(31) = 0.39 \pm 0.08$, $P = 0.022$) children, as well as at T2 for relatively younger ($r(31) = 0.64 \pm 0.07$, $P = 0.0001$) and older ($r(31) = 0.51 \pm 0.07$, $P = 0.005$) children.

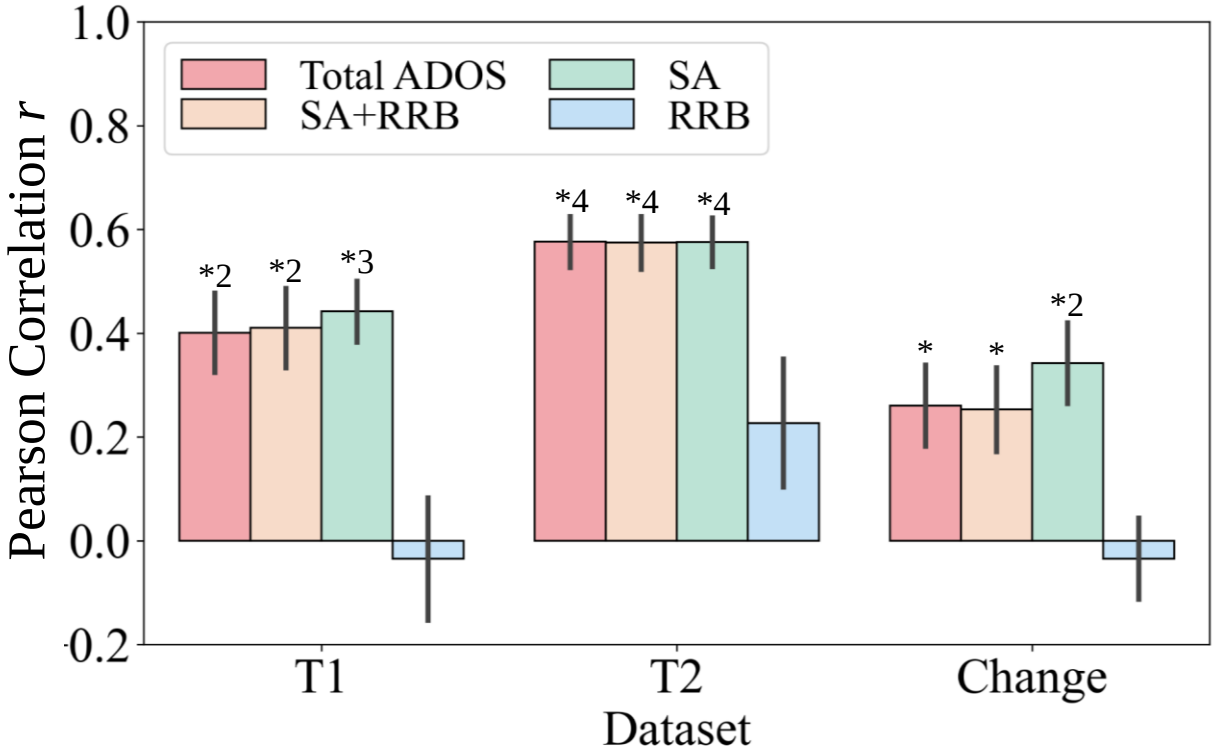

Fig. 3: Accuracy of the ASE algorithms. Pearson correlation coefficients computed between the estimated ADOS-2 scores and actual scores for T1 (left), T2 (middle), and their change across time-points (right). **Red:** Estimated total ADOS-2 scores versus actual total ADOS-2 scores. **Orange:** Sum of separately estimated SA and RRB scores versus actual total ADOS-2 scores. **Green:** Estimated ADOS-2 SA scores versus actual ADOS-2 SA scores. **Blue:** Estimated ADOS-2 RRB scores versus actual ADOS-2 RRB scores. **Error bars:** standard deviation across 500 testing iterations. **Asterisks:** statistical significance determined by permutation test (*: P value < 0.05, *2: P value ≤ 0.01, *2: P value ≤ 0.001, ∗4: P value ≤ 0.0001).

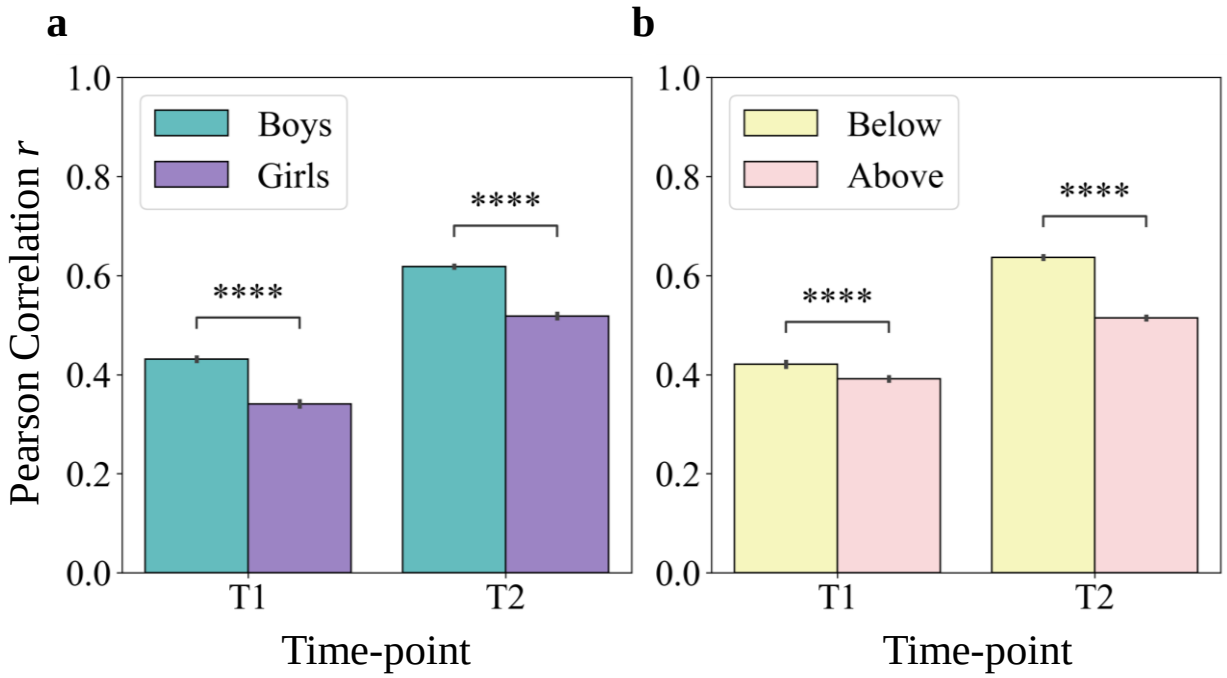

Fig. 4: Comparison of total ADOS score ASE algorithm accuracy. (**a**) Boys and girls. (**b**) Younger and older children. Accuracy was significantly higher for boys (n=43, **green**) versus girls (n=19, **purple**) and for children below (**yellow**) versus above (**light pink**) the median age. **Error bars**: standard error of the mean across 500 testing iterations. **Asterisks**: significant difference across groups (*t*-test, ****: *P* value < 0.001).

Discussion

In the current study, a CNN model trained on speech recordings of 146 ASD children was able to partially predict changes in the social symptom severity (i.e., ADOS-2 SA scores) of 62 entirely unrelated ASD children. This demonstrates that the model was able to identify speech characteristics that were indicative of social ASD symptoms in one group of children and successfully generalize their use to estimate changes in the severity of social symptoms in another group of children. For this to work, the model had to accurately predict the ADOS-2 SA scores of the children in each of two separate assessments separated by 1-2 years. Note that we did not train the model to identify changes in ADOS-2 SA scores, the non-trivial ability to predict such changes was the outcome of the model's ability to separate speech characteristics associated with the child's identity (which were identical across time-points) and speech characteristics associated with the severity of their ASD social symptoms (which changed across time-points in some of the children – Fig. 2).

It is remarkable that such prediction was possible at all given the large heterogeneity across ASD children who can manifest their social difficulties in dramatically different behaviors that are not necessarily related to their speech (e.g., difficulties with eye contact, imitation, joint attention, and other social behaviors [3], [34]). Nevertheless, 20-33% of the variance in ADOS-2 SA scores at each time-point and 11.5% of the variance in score changes across time-points were explained by the model (Fig. 3). We speculate that enlarging the training dataset could increase the accuracy

of the model even further but expect that achieving higher levels of prediction would require integration with additional types of behavioral data such as eye tracking [35] and movement analysis [36] from the same children.

In contrast to the significant prediction abilities of social symptoms in the tested sample, the model performed poorly in predicting RRB symptom severity (Fig. 3). We believe there may be several reasons for this. First, the limited range of the ADOS-2 RRB scale (0-8) relative to the SA scale (0-22) may make it difficult for the CNN model to identify differences across children. Indeed, a recent study reported that the limited number of items on the RRB scale resulted in poor scale reliability across participants [37]. Second, the selected speech features in the current study exhibited weaker correlations with RRB than SA scores (Fig. 1). Extraction of other speech features, such as phrase or intonation repetitions (indicative of echolalia) may improve the accuracy of ADOS-2 RRB score estimates. Regardless, our results motivate separate modeling of social and RRB symptom domains as each of them is likely associated with distinct features of speech and changes independently of the other (Fig. 2**d**). Note that successful prediction of total ADOS-2 scores (Fig. 3) was likely due to the fact that SA scores account for two-thirds of the total scores.

Finally, our results suggest that the model's prediction accuracy was significantly better for boys than girls. Since our training sample included 117 boys and 29 girls and our testing sample included 43 boys and 19 girls, it may be the case that better prediction for boys was driven by the availability of more exemplars for this group. Nevertheless, it is remarkable that prediction accuracy was still relatively high for the girls, with 12-27% of the variance in the test dataset explained at each of the two ADOS-2 assessments (Fig. 4**a**). This suggests that the model was able to generalize predictions using the same speech characteristics across individuals of both sexes. Training with larger samples of ASD girls will likely further improve the model's accuracy. Prediction accuracy was also significantly higher for the younger relative to the older children in our sample (Fig. 4), suggesting a stronger relationship between speech characteristics and ASD severity at earlier ages. However, prediction was still relatively high for older children, with 15-26% of the variance in the test dataset explained at each of the two ADOS-2 assessments (Fig. 4**b**). This suggests that the model was able to generalize predictions using the same speech characteristics across ASD children of different ages.

Speech characteristics differ across TD and ASD individuals

A variety of studies have reported that individuals with ASD, on average, speak differently than TD individuals [4], [8]–[16]. According to these studies, ASD individuals exhibit atypical speech characteristics, including significantly fewer speech sounds [13], fewer phonemes per utterance [11], and fewer conversational turns [13] as well as significantly larger mean pitch [9], [19], larger pitch range [8], [9], and larger pitch variability [8], [9] than TD children. Differences in these and other speech characteristics have enabled the development of classification algorithms that can separate ASD and TD individuals with reported accuracy rates as high as 75-98% [17]–[23]. However, it is important to note that most classification studies have included small samples with <40 ASD participants per study who were recorded at different ages and in different contexts ranging from 12-minute parent-child free play sessions [17] to repetition of non-word syllable groups [23]. Identifying the specific age-dependent speech characteristics that reliably distinguish

between ASD and TD children will require additional validation with considerably larger datasets that capture broad ASD heterogeneity.

Quantifying ASD severity according to speech recordings

In parallel with the efforts described above to classify ASD and TD children based on their speech, several studies have attempted to quantify ASD severity based on speech characteristics. The feasibility of this approach is motivated by results presented in the current and previous study from our lab [28] demonstrating that multiple speech features were significantly correlated with ADOS-2 scores (Fig. 1 and Supplementary Table S1). Moreover, specific features were correlated with ADOS-2 SA scores, while others were correlated with ADOS-2 RRB scores, suggesting that unique speech features contribute to quantifying the two distinct core symptoms of ASD.

While significant, the univariate correlations described above were of medium effect sizes or less ($r<0.5$). Utilizing multivariate approaches could potentially identify more robust relationships between multiple speech features and ADOS-2 scores. Hence, three recent studies [28]–[30] have attempted to predict ADOS-2 scores from speech recordings of ASD individuals using multivariate techniques. In the first [29], a synthetic random forest model was trained to estimate ADOS-2 SA scores using vocalization rates and conversational features extracted from ADOS-2 recordings of 33 ASD children. The algorithm was able to predict ADOS-2 SA scores that were significantly correlated with actual scores ($r = 0.634$). The second study utilized a DNN model to estimate four ADOS-2 item scores using hundreds of conversational and acoustic features extracted from ADOS-2 recordings of 88 high-functioning ASD adolescents/adults [30]. This algorithm was able to estimate scores that exhibited significant Spearman correlations with the actual scores (ρ = 0.519-0.645). Finally, in a previous study from our lab [28], we demonstrated that a CNN model was able to estimate ADOS-2 total scores that were significantly correlated with actual scores ($r = 0.718$) when using 60 conversational and acoustic features extracted from ADOS-2 recordings of 72 children, 56 of them with ASD.

The current study extends previous work in several critical ways. First, the current study utilized the largest dataset available to date (270 ADOS-2 recordings), which was at least three times larger than those used by previous studies, with 72 [28], 33 [29], and 88 [30] recordings. Note that speech segments in all recordings were manually annotated to ensure accuracy. Second, given the large number of participating children, we were able to test algorithm accuracy on a much larger sample of children (n=62) who were recorded twice during two ADOS-2 assessments separated by 1-2 years. This enabled us to test the reliability of the algorithms across two separate time-points. Third, having data from two time-points also enabled us to examine the accuracy of the algorithms in accurately detecting changes over time – a critical feature that may enable future use as an outcome measure for clinical trials assessing treatment efficacy. Fourth, in this study, we examined the ability to separate the two core ASD symptoms by training separate algorithms to estimate ADOS-2 SA and ADOS-2 RRB symptoms, respectively. The results demonstrated that this separation is indeed important and that different speech features are likely to contribute to the estimation of each symptoms domain. Fifth, the larger sample size enabled us to demonstrate that algorithm accuracy was significantly higher in boys and younger children at both time-points. This suggests that algorithms may require specific additional training to achieve equivalent accuracy across different subgroups of ASD participants. Sixth, the recordings utilized in the current study were performed over a 6-year period in a busy public healthcare medical center that services a

population of ~1 million people. Recordings were performed with a wall mounted microphone (see Methods) in "real world" acoustic conditions (i.e., variable environmental noises such as air conditioning). This further demonstrates the robustness of the algorithms such that reliable estimates were possible despite variability in recording conditions.

While the accuracy of the algorithms presented in this study may be somewhat lower than those reported in previous studies, is likely to be more realistic of current abilities in an active clinical setting. We believe that the encouraging results in the current study motivate investment in creating larger datasets that are likely to yield higher accuracy while establishing stronger robustness across time-points and higher sensitivity to change.

Limitations

The current study had several limitations. First, we did not examine the language content of recorded children. We did not transcribe the recordings and did not apply any analyses related to language structure or vocabulary, which is likely to improve the estimation of ASD severity [4], [24]. Second, we did not identify echolalia events that are likely to be informative of RRB symptoms. Indeed, our poor results estimating RRB scores suggest that different speech features are necessary for estimating severity in this domain. Third, we did not systematically examine the impact of language and cognitive abilities on algorithm estimates. Future studies may want to train separate algorithms for estimating cognitive and language scores with the goal of separately modeling multiple behavioral domains (analogous to our separation of the SA and RRB scores). Note that this approach would likely require a considerably larger dataset of participants where these features are not correlated. Finally, future studies may benefit from applying noise reduction and/or multi-speaker analysis techniques to improve the quality of analyzed vocal segments.

Conclusions

Speech recordings contain revealing information about the social symptom severity of ASD children. This study and others demonstrate that it is possible to partially estimate ADOS-2 SA scores by utilizing automated speech analysis techniques. The current study extends previous knowledge by demonstrating high robustness of these estimates across two time-points as well as the ability to detect changes across time-points even when recordings are performed in "real life" settings within a busy community healthcare center. Large freely accessible repositories of ADOS-2 recordings are needed to further improve the development of such algorithms, which are likely to be highly useful phenotyping tools for clinical and basic ASD research.

Methods

Overview. Our speech-based ASE system (Fig. 5) utilized a dataset of 270 recordings where child speech segments were manually annotated. The system performed the following main steps: (1) separating child segments into individual child vocalizations; (2) extracting acoustic and conversational features from the child vocalizations; (3) estimation of ASD severity (ADOS-2 scores) from matrix of speech features per child.

The ASE model was trained using 146 recordings of children who completed a single assessment (i.e., training phase) and system performance was determined using independent

recordings from 62 children who completed two separate ADOS-2 recordings. In the subsequent sections, we will describe the specifics of these steps.

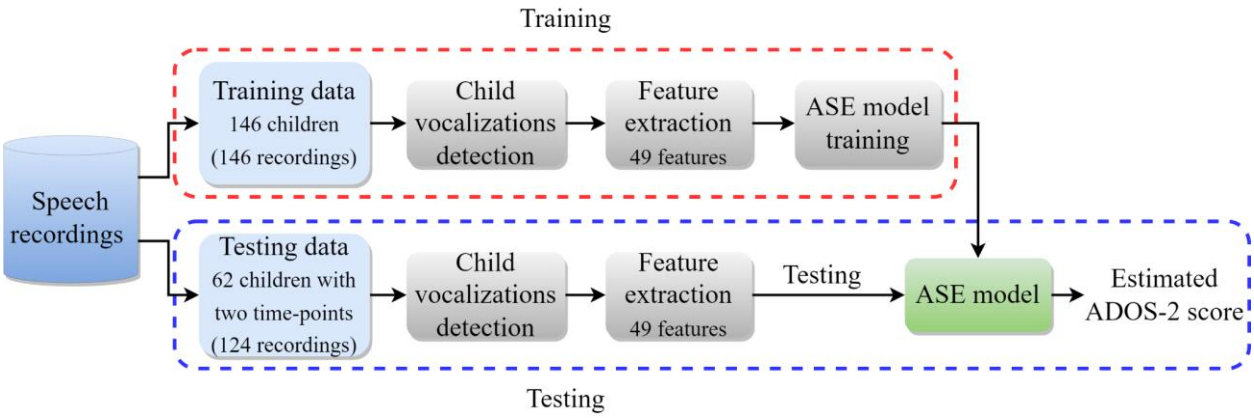

Fig. 5: Block diagram of the speech-based ASE system. Training (top): Data selection and processing as part of ASE model training. Testing (bottom): Data selection, processing, and estimation of autism severity score from speech recordings. ASE: Autism Severity Estimation.

Participants and setting. This study analyzed data obtained from the Azrieli National Centre for Autism and Neurodevelopment Research (ANCAN), a collaboration between Ben Gurion University of the Negev (BGU) and ten partner clinical centers where ASD is diagnosed throughout Israel. ANCAN manages the national autism database of Israel with data from >3000 children in 2023 and growing ([38], [39]). All recordings used in the current study were performed in a single ANCAN assessment room located at Soroka University Medical Center (SUMC), the largest partner clinical site. A total of 208 children (1-7-years-old) who completed at least one ADOS-2 assessment between 2015 and 2021 and received an ASD diagnosis were included in this study. Of the participating children, 146 completed a single ADOS-2 assessment and 62 completed two ADOS-2 assessments at two time-points separated by 10-29 months, yielding 270 ADOS-2 assessments in total (Table 1). ADOS-2 is a semi-structured assessment where a clinician administers specific tasks, observes the behavior of the child, and scores their behavior. The total ADOS-2 severity score is in the range of 0-30 and is the sum of Social Affect (SA, 0-22) and Restricted and Repetitive Behavior (RRB, 0-8) scores, with higher scores indicating more severe symptoms. All ADOS-2 assessments were performed by a clinician with research reliability. In addition, all participating children had ASD diagnoses that were confirmed by both a developmental psychologist and either a child psychiatrist or a pediatric neurologist, according to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria [31]. Informed consent was obtained from all parents, and the study received approval from the SUMC Helsinki committee.

Recording setup. All recordings were performed during ADOS-2 assessments using a single microphone (CHM99, AKG, Vienna) located on a wall, ~1-2m from the child, and connected to a sound card (US-16x08, TASCAM, California). Each ADOS-2 session lasted ~40-minutes (40.87 ± 11.91 min) and was recorded at a sampling rate of 44.1 kHz, 16 bits/sample (down-sampled to 16 kHz).

Table 1: Participating children's characteristics

	Single assessment	Two assessments	
	(*N* = 146)	T1 (*N* = 62)	T2 (*N* = 62)
Mean (SD)			
Age (years)	4.29 (1.35)	3.67 (0.97)	4.96 (0.94)
ADOS-2 Total	14.90 (5.87)	14.77 (5.95)	14.97 (5.58)
ADOS-2 SA	10.74 (4.95)	11.15 (5.17)	10.63 (4.53)
ADOS-2 RRB	4.16 (1.88)	3.63 (1.72)	4.34 (1.85)
N (%)			
Sex			
Male	117 (80)	43 (69)	43 (69)
Female	29 (20)	19 (31)	19 (31)
Module			
Module T	19 (13)	9 (15)	0 (0)
Module 1	46 (32)	30 (48)	29 (47)
Module 2	46 (32)	21 (34)	19 (31)
Module 3	35 (23)	2 (3)	14 (22)

Detection of child vocalizations. Audio recordings were manually divided and labeled as child, therapist, parent, simultaneous speech (i.e., speech of more than one speaker), or noise (e.g., movements in the room) segments. All remaining segments were automatically labeled as silence. Only child-labeled segments were used in further analysis. These segments included speech, laughing, moaning, crying, and screaming. The child segments often contained multiple vocalizations (e.g., multiple utterances) separated by silence. We separated each segment into multiple vocalizations using energy thresholds of 2.79dB and 0.4dB above the background noise level to define the beginning and the end, respectively, of each vocalization [28]. Vocalizations that were shorter than 110ms were excluded from further analysis (too short to contain an utterance). An example of manual segmentation and detection of vocalizations is shown in Supplementary Fig. S3.

Features. We extracted 49 features that were categorized into nine groups: pitch, formants, jitter, voicing, energy, Zero-Crossing Rate (ZCR), spectral slope, duration, and quantity/number of vocalizations. All features, except duration and quantity, were first extracted in 40ms windows (window overlap of 75%), resulting in a vector of feature values per vocalization. The minimum,

the maximum, and the mean pitch of the voiced vocalizations (across windows) were computed, deriving one value for each vocalization. We then selected a group of 10 consecutive vocalizations and computed the mean and variance across vocalizations for relevant features (Table 2). We also computed the mean duration of vocalizations and the overall number of vocalizations in the recording. Taken together, these steps yielded a vector with 49 values corresponding to the 49 features per 10 vocalizations. We performed this procedure 100 times, selecting random groups of ten consecutive vocalizations from the recording. Combining these 100 samples yielded a features matrix of 100×49 per child (Supplementary Fig. S4), with the last column (quantity of vocalizations) containing the same value across all rows. Features included:

Frequency related features:

- *Pitch (F0):* Vocal cords vibration frequency (the fundamental frequency) that exists only in voiced speech (e.g., vowels). Voiced Vocalization (VV) was defined as a vocalization where most of its frames (≥60%) [10] were voiced (voicing threshold 0.45).
- *Formants:* The resonant frequencies of the vocal tract that shape vowel sounds [40]. The first two formants (F1 and F2) relate to tongue position (vertical and horizontal) and influence vowel quality. Their bandwidths affect the clarity of speech.
- *Jitter:* Variation across adjacent pitch values representing frequency instability [41].
- *Voicing*: Pitch peak amplitude as determined by the autocorrelation function.

Pitch and formants were calculated using the PRAAT software [42], with a pitch range set to 60–1600 Hz (a wide range to increase sensitivity to atypical vocal characteristics).

Energy/amplitude related features:

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

Spectral features:

- *Zero-Crossing Rate (ZCR):* The number of zero-crossings apparent in audio segments with child vocalizations [43].
- *Spectral slope:* The slope of the linear regression on the logarithmic power spectrum within the frequency bands of 20-500 Hz (lower band) and 500-1500 Hz (higher band) [44], [45].

Conversational features:

- *Duration:* Child's mean vocalization length.
- *Quantity*: The total number of vocalizations.

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

Table 2: Acoustic and conversational features extracted from child vocalizations.

#	Feature type	Feature
1	Pitch (F0)	Mean F0 across consecutive† vocalizations
2		Variance of F0 across consecutive vocalizations
3		Correlation of coefficient (σ^2/μ) across consecutive vocalizations
4		Mean of minimum F0: the minimum F0 calculated across vocalization frames and then averaged across consecutive vocalizations
5		Variance of minimum F0: the minimum F0 calculated across vocalization frames. Their variance was computed across the minimum F0 values of consecutive vocalizations
6		Mean of maximum F0: the maximum F0 calculated across vocalization frames and then averaged across consecutive vocalizations
7		Variance of maximum F0: the maximum F0 calculated across vocalization frames. Their variance was computed across the maximum F0 values of consecutive vocalizations
8		Mean F0 across VVs within the consecutive vocalizations
9		Variance of F0 across VVs within the consecutive vocalizations
10		Variance of the mean F0 across VVs: the mean F0 calculated for each vocalization. Their variance was computed across the VVs within the consecutive vocalizations
11	Formants (F) & Bandwidth (BW)	Mean F1 across consecutive vocalizations
12		Variance of F1 across consecutive vocalizations
13		Mean F2 across consecutive vocalizations
14		Variance of F2 across consecutive vocalizations
15		Mean |F1-F2| across consecutive vocalizations
16		Variance of |F1-F2| across consecutive vocalizations
17		Mean BW1 across consecutive vocalizations
18		Variance of BW1 across consecutive vocalizations
19		Mean BW2 across consecutive vocalizations
20		Variance of BW2 across consecutive vocalizations
21	Jitter	Mean jitter across consecutive vocalizations
22		Variance of jitter across consecutive vocalizations
23	Voicing	Mean voicing across consecutive vocalizations
24		Variance of voicing across consecutive vocalizations
25	Energy (E)	Mean E across consecutive vocalizations
26		Mean ΔE across consecutive vocalizations
27		Mean ΔΔE across consecutive vocalizations
28		Mean |ΔE| across consecutive vocalizations
29		Variance E across consecutive vocalizations
30		Variance of ΔE across consecutive vocalizations
31		Variance of ΔΔE across consecutive vocalizations

32		Variance of |ΔE| across consecutive vocalizations
33	Zero-Crossing Rate (ZCR)	Mean ZCR across consecutive vocalizations
34		Variance of ZCR consecutive vocalizations
35		Mean ZCR across VVs within consecutive vocalizations
36		Variance of ZCR across VVs within the consecutive vocalizations
37		Mean ZCR across UVs within the consecutive vocalizations
38		Variance of ZCR across UVs within the consecutive vocalizations
39	Spectral slope (SpSl)	Mean SpSl [range: 20-500 Hz] Hz across consecutive vocalizations
40		Variance of SpSl [range: 20-500 Hz] across consecutive vocalizations
41		Mean SpSl [range: 50-1500 Hz] across consecutive vocalizations
42		Variance of SpSl [range: 50-1500 Hz] across consecutive vocalizations
43		Mean SpSl [range: 20-500 Hz] across VVs within consecutive vocalizations
44		Mean SpSl [range: 20-500 Hz]across UVs within consecutive vocalizations
45		Mean SpSl [range: 50-1500 Hz] across VVs within consecutive vocalizations
46		Mean SpSl [range: 50-1500 Hz] across UVs within consecutive vocalizations
47	Duration	Mean duration across consecutive vocalizations
48		Variance of duration across consecutive vocalizations
49	Quantity	Total number of vocalizations

† The number of consecutive vocalizations was set to ten.
VV – Voiced Vocalization, UV – Unvoiced Vocalization.

Training the Autism Severity Estimation (ASE) models. Training was performed with data from the 146 children who completed a single ADOS-2 session only. Feature matrices were used to train three deep learning models with an identical CNN architecture (Fig. 6 and Fig. 7). The first model estimated ADOS-2 SA scores, the second estimated ADOS-2 RRB scores, and the third estimated total ADOS-2 scores (sum of SA and RRB scores). Training was based on minimizing the Mean Squares Error (MSE) of a regression analysis between estimated and actual ADOS-2 scores. We used the RMSprop (Root Mean Square Propagation) as the optimization algorithm [46], and we applied the random search algorithm to tune the batch size, number of epochs, and the learning rate using a stratified 5-fold cross-validation method on the training dataset. Early stopping of patience 20 epochs was applied to reduce overfitting [47].

Testing the ASE system. To test the performance of each model, we estimated the ADOS-2 scores from an independent sample of 62 children who completed two ADOS-2 assessments at T1 and T2. We estimated ADOS-2 values separately at each time-point and then computed the difference across time-points (T2-T1) to estimate the change in ADOS-2 scores. In this scheme, a positive value (higher value at T1) corresponds to a decrease in symptoms severity (i.e., improvement), while a negative value corresponds to an increase in symptom severity.

The performance of each model was measured using the Pearson correlation, calculated between estimated and actual ADOS-2 scores of the testing data. The mean and the standard deviation of the Pearson correlations were calculated iteratively using 500 different models (500 iterations). In each iteration, the model was trained using 80% of the 146 recordings using random data split while ensuring the balancing of ADOS-2 scores, following a methodology similar to [28]. In total, we had 50 different data splits, each including 10 feature matrices that were generated by randomly selecting groups of ten consecutive vocalizations (50×10=500 iterations).

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

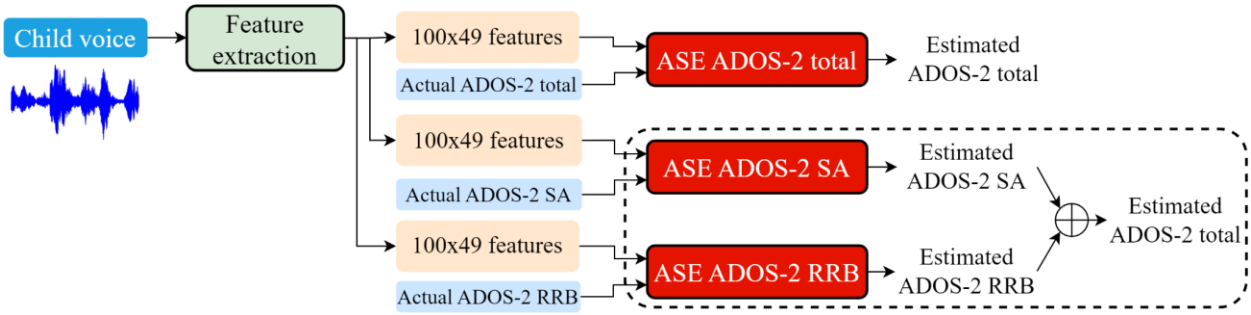

Fig. 6: Three autism severity estimation models. The ASE ADOS-2 total model was trained to estimate total ADOS-2 scores, while the ASE ADOS-2 SA and ASE ADOS-2 RRB were trained to estimate SA and RRB scores, respectively. Estimated SA and RRB scores were then summed to yield a second estimated ADOS-2 total score.

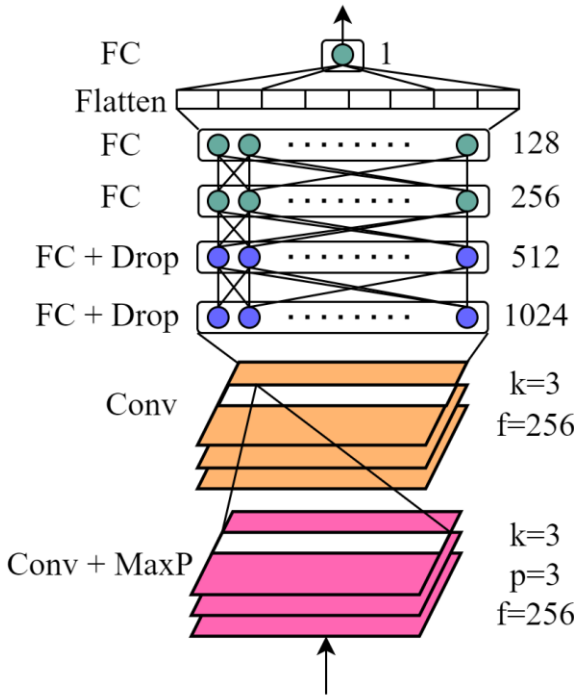

Fig. 7: The CNN architecture of the ASE models. Conv: for 1d convolution layer, Conv + MaxP: convolution layer followed by max pooling layer, FC: fully connected layer, FC + Drop: fully connected layer followed by dropout of 0.5, k: kernel size, p: pooling size, f: filters.

Statistical Analysis. All statistics were performed with custom-written code in Python. The statistical significance of differences across groups (e.g., boys and girls) was estimated using an independent *t*-test using the Statannotations software package [48]. The associations between speech features and ADOS-2 scores were determined using Pearson correlation coefficient analysis. To evaluate their statistical significance, we performed a random permutation test where we randomly shuffled the actual ADOS-2 scores across children and calculated a correlation between the predicted scores by the model and the shuffled scores (R_{perm}). This procedure was performed 1000 times with different shuffles on each iteration, thereby creating a null distribution of random correlation coefficients, which retains the distribution of the data that is not necessarily normal. For a correlation to be considered significant, the actual correlation (without shuffling) had to exceed the 97.5 percentile of the null distribution.

References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *Lancet*, vol. 392, no. 10146, pp. 508–520, Aug. 2018, doi: 10.1016/S0140-6736(18)31129-2.

[2] N. Gabbay-Dizdar *et al.*, "Early diagnosis of autism in the community is associated with marked improvement in social symptoms within 1–2 years," *Autism*, p. 136236132110490, Oct. 2021, doi: 10.1177/13623613211049011.

[3] H. Tager-Flusberg and C. Kasari, "Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum," *Autism Res.*, vol. 6, no. 6, pp. 468–478, Dec. 2013, doi: 10.1002/aur.1329.

[4] A. C. Salem *et al.*, "Evaluating atypical language in autism using automated language measures," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/S41598-021-90304-5.

[5] N. A. Chi *et al.*, "Classifying Autism From Crowdsourced Semi-structured Speech Recordings: Machine Learning Model Comparison Study," *JMIR Pediatr. Parent.*, vol. 5, no. 2, p. e35406, Apr. 2022, doi: 10.2196/35406.

[6] J. Mostek, "Cognitive Development and Language Acquisition in Autistic Children," *Sci. Insights*, vol. 41, no. 6, pp. 719–724, Nov. 2022, doi: 10.15354/si.22.re091.

[7] S. Loukusa, "Autism Spectrum Disorder," in *Handbook of Pragmatic Language Disorders*, L. Cummings, Ed. Cham: Springer International Publishing, 2021, pp. 45–78.

[8] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal Speech Spectrum and Increased Pitch Variability in Young Autistic Children," *Front. Hum. Neurosci.*, vol. 4, p. 237, Jan. 2011, doi: 10.3389/fnhum.2010.00237.

[9] S. Z. Asghari, S. Farashi, S. Bashirian, and E. Jenabi, "Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study," *Sci. Rep.*, vol. 11, no. 1, p. 23093, Nov. 2021, doi: 10.1038/s41598-021-02487-6.

[10] D. K. Oller *et al.*, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proc. Natl. Acad. Sci.*, vol. 107, no. 30, pp. 13354–13359, Jul. 2010, doi: 10.1073/pnas.1003882107.

[11] J. M. Moffitt *et al.*, "Objective measurement of vocalizations in the assessment of autism spectrum disorder symptoms in preschool age children," *Autism Res.*, pp. 1–10, Apr. 2022, doi: 10.1002/aur.2731.

[12] A. S. Warlaumont, J. A. Richards, J. Gilkerson, and D. K. Oller, "A Social Feedback Loop for Speech Development and Its Reduction in Autism," *Psychol. Sci.*, vol. 25, no. 7, pp.

1314–1324, Jul. 2014, doi: 10.1177/0956797614531023.

[13] S. F. Warren *et al.*, “What Automated Vocal Analysis Reveals About the Vocal Production and Language Learning Environment of Young Children with Autism,” *J. Autism Dev. Disord.*, vol. 40, no. 5, pp. 555–569, May 2010, doi: 10.1007/s10803-009-0902-5.

[14] E. F. Ferguson, A. S. Nahmias, S. Crabbe, T. Liu, D. S. Mandell, and J. Parish-Morris, “Social language opportunities for preschoolers with autism: Insights from audio recordings in urban classrooms,” *Autism*, vol. 24, no. 5, pp. 1232–1245, Jul. 2020, doi: 10.1177/1362361319894835.

[15] A. M. Plumb and A. M. Wetherby, “Vocalization Development in Toddlers With Autism Spectrum Disorder,” *J. Speech, Lang. Hear. Res.*, vol. 56, no. 2, pp. 721–734, Apr. 2013, doi: 10.1044/1092-4388(2012/11-0104).

[16] E. J. Tenenbaum *et al.*, “A Six-Minute Measure of Vocalizations in Toddlers with Autism Spectrum Disorder,” *Autism Res.*, vol. 13, no. 8, pp. 1373–1382, Aug. 2020, doi: 10.1002/aur.2293.

[17] F. B. Pokorny *et al.*, “Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, no. August, pp. 309–313, 2017, doi: 10.21437/Interspeech.2017-1007.

[18] A. Mohanta and V. K. Mittal, “Classifying Speech of ASD Affected and Normal Children Using Acoustic Features,” in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6, doi: 10.1109/NCC48643.2020.9056084.

[19] A. Mohanta and V. K. Mittal, “Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features,” *Comput. Speech Lang.*, vol. 72, p. 101287, Mar. 2022, doi: 10.1016/j.csl.2021.101287.

[20] M. Asgari and I. Shafran, “Robust and Accurate Features for Detecting and Diagnosing Autism Spectrum Disorders,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. January, pp. 191–194, 2013.

[21] L. D. Yankowitz, R. T. Schultz, and J. Parish-Morris, “Pre- and Paralinguistic Vocal Production in ASD: Birth Through School Age,” *Curr. Psychiatry Rep.*, vol. 21, no. 12, p. 126, Dec. 2019, doi: 10.1007/s11920-019-1113-1.

[22] S. Lee, E. J. Yeo, S. Kim, and M. Chung, “Knowledge-driven speech features for detection of Korean-speaking children with autism spectrum disorder*,” *Phonetics Speech Sci.*, vol. 15, no. 2, pp. 53–59, Jun. 2023, doi: 10.13064/KSSS.2023.15.2.053.

[23] F. Briend, C. David, S. Silleresi, J. Malvy, S. Ferré, and M. Latinus, “Voice acoustics allow

classifying autism spectrum disorder with high accuracy," *Transl. Psychiatry*, vol. 13, no. 1, p. 250, Jul. 2023, doi: 10.1038/s41398-023-02554-8.

[24] H. MacFarlane, A. C. Salem, L. Chen, M. Asgari, and E. Fombonne, "Combining voice and language features improves automated autism detection," *Autism Res.*, vol. 15, no. 7, pp. 1288–1300, Jul. 2022, doi: 10.1002/AUR.2733.

[25] B. B. Maddox *et al.*, "The Accuracy of the ADOS-2 in Identifying Autism among Adults with Complex Psychiatric Conditions," *J. Autism Dev. Disord.*, vol. 47, no. 9, p. 2703, Sep. 2017, doi: 10.1007/S10803-017-3188-Z.

[26] 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.*, vol. 64, no. 5, May 2023, doi: 10.1111/JCPP.13802.

[27] C. Klaiman *et al.*, "Expert Clinician Certainty in Diagnosing Autism Spectrum Disorder in 16-30-Month-Olds: A Multi-site Trial Secondary Analysis," *J. Autism Dev. Disord.*, 2022, doi: 10.1007/S10803-022-05812-8.

[28] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network," *IEEE Access*, vol. 8, pp. 139489–139500, 2020, doi: 10.1109/ACCESS.2020.3012532.

[29] S. Sadiq, M. Castellanos, J. Moffitt, M. Shyu, L. Perry, and D. Messinger, "Deep Learning Based Multimedia Data Mining for Autism Spectrum Disorder (ASD) Diagnosis," in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 847–854, doi: 10.1109/ICDMW.2019.00124.

[30] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Learning Converse-Level Multimodal Embedding to Assess Social Deficit Severity for Autism Spectrum Disorder," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, vol. 2020-July, pp. 1–6, doi: 10.1109/ICME46284.2020.9102869.

[31] D. Trembath *et al.*, "Profiles of vocalization change in children with autism receiving early intervention," *Autism Res.*, vol. 12, no. 5, pp. 830–842, May 2019, doi: 10.1002/aur.2075.

[32] E. Waizbard-Bartov *et al.*, "Trajectories of Autism Symptom Severity Change During Early Childhood," *J. Autism Dev. Disord.*, vol. 51, no. 1, pp. 227–242, Jan. 2021, doi: 10.1007/s10803-020-04526-z.

[33] E. Ferrari, "Artificial Intelligence for Autism Spectrum Disorders," in *Artificial Intelligence in Medicine*, Cham: Springer International Publishing, 2022, pp. 1579–1593.

[34] G. Vivanti, K. Bottema-Beutel, and L. Turner-Brown, "Understanding and Addressing Restricted and Repetitive Behaviors in Children with Autism," in *Clinical Guide to Early*

Interventions for Children with Autism, 2020, pp. 61–77.

[35] I. Avni *et al.*, “Children with autism observe social interactions in an idiosyncratic manner,” *Autism Res.*, vol. 13, no. 6, pp. 935–946, Jun. 2020, doi: 10.1002/aur.2234.

[36] I. Budman *et al.*, “Quantifying the social symptoms of autism using motion capture,” *Sci. Rep.*, vol. 9, no. 1, p. 7712, Dec. 2019, doi: 10.1038/s41598-019-44180-9.

[37] T. W. Frazier *et al.*, “Reliability of the Commonly Used and Newly-Developed Autism Measures,” *J. Autism Dev. Disord.*, Apr. 2023, doi: 10.1007/s10803-023-05967-y.

[38] I. Dinstein *et al.*, “The National Autism Database of Israel: a Resource for Studying Autism Risk Factors, Biomarkers, Outcome Measures, and Treatment Efficacy,” *J. Mol. Neurosci.*, vol. 70, no. 9, pp. 1303–1312, Sep. 2020, doi: 10.1007/s12031-020-01671-z.

[39] G. Meiri *et al.*, “The Negev Hospital-University-Based (HUB) Autism Database,” *J. Autism Dev. Disord.*, vol. 47, no. 9, pp. 2918–2926, 2017, doi: 10.1007/s10803-017-3207-0.

[40] P. Padmini, D. Gupta, M. Zakariah, Y. A. Alotaibi, and K. Bhowmick, “A Simple Speech Production System Based on Formant Estimation of a Tongue Articulatory System Using Human Tongue Orientation,” *IEEE Access*, vol. 9, pp. 4688–4710, 2021, doi: 10.1109/ACCESS.2020.3048076.

[41] J. Rusz *et al.*, “Characteristics of motor speech phenotypes in multiple sclerosis,” *Mult. Scler. Relat. Disord.*, vol. 19, pp. 62–69, Jan. 2018, doi: 10.1016/j.msard.2017.11.007.

[42] P. Boersma and V. van Heuven, “Speak and unSpeak with Praat,” *Glot Int.*, vol. 5, no. 9–10, pp. 341–347, 2001.

[43] L. G Pillai and E. Sherly, “A Deep Learning Based Evaluation of Articulation Disorder and Learning Assistive System for Autistic Children,” *Int. J. Nat. Lang. Comput.*, vol. 6, no. 5, pp. 19–36, 2017, doi: 10.5121/ijnlc.2017.6502.

[44] L. Tamarit, M. Goudbeek, and K. Scherer, “Spectral Slope Measurements in Emotionally Expressive Speech,” in *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery*, 2008, pp. 1–4, paper 007.

[45] F. Eyben *et al.*, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016, doi: 10.1109/TAFFC.2015.2457417.

[46] A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova, “The Impact of Multi-Optimizers and Data Augmentation on TensorFlow Convolutional Neural Network Performance,” *Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, vol. April, pp. 140–145, 2018, doi: 10.1109/MIPR.2018.00032.

[47] L. Prechelt, “Early Stopping — But When?,” in *Neural Networks: Tricks of the Trade*, vol. 134, no. 1, 2012, pp. 53–67.

[48] F. C. And *et al.*, “Statannotations.” Zenodo, 2022, doi: 10.5281/zenodo.7213391.

Acknowledgments

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