HHS Public Access

Author manuscript

Autism Res. Author manuscript; available in PMC 2022 July 05.

Published in final edited form as:

Autism Res. 2022 July ; 15(7): 1288–1300. doi:10.1002/aur.2733.

Combining voice and language features improves automated autism detection

Heather MacFarlane[1], **Alexandra C. Salem**[1], **Liu Chen**[2,3], **Meysam Asgari**[2,3], **Eric Fombonne**[1,2,3]

[1]Department of Psychiatry, Oregon Health & Science University, Portland, Oregon, USA

[2]Institute on Development and Disability, Oregon Health & Science University, Portland, Oregon, USA

[3]Department of Pediatrics, Oregon Health & Science University, Portland, Oregon, USA

Abstract

Variability in expressive and receptive language, difficulty with pragmatic language, and prosodic difficulties are all features of autism spectrum disorder (ASD). Quantifying language and voice characteristics is an important step for measuring outcomes for autistic people, yet clinical measurement is cumbersome and costly. Using natural language processing (NLP) methods and a harmonic model of speech, we analyzed language transcripts and audio recordings to automatically classify individuals as ASD or non-ASD. One-hundred fifty-eight participants (88 ASD, 70 non-ASD) ages 7 to 17 were evaluated with the autism diagnostic observation schedule (ADOS-2), module 3. The ADOS-2 was transcribed following modified SALT guidelines. Seven automated language measures (ALMs) and 10 automated voice measures (AVMs) for each participant were generated from the transcripts and audio of one ADOS-2 task. The measures were analyzed using support vector machine (SVM; a binary classifier) and receiver operating characteristic (ROC). The AVM model resulted in an ROC area under the curve (AUC) of 0.7800, the ALM model an AUC of 0.8748, and the combined model a significantly improved AUC of 0.9205. The ALM model better detected ASD participants who were younger and had lower language skills and shorter activity time. ASD participants detected by the AVM model had better language profiles than those detected by the language model. In combination, automated measurement of language and voice characteristics successfully differentiated children with and without autism. This methodology could help design robust outcome measures for future research.

Lay Summary:

Correspondence Heather MacFarlane, Department of Psychiatry, Oregon Health & Science University, Portland, OR, USA. macfarlh@ohsu.edu.
Heather MacFarlane and Alexandra C. Salem contributed equally to this work.

ETHICS STATEMENT

This study was reviewed and approved by the Institutional Review Board of OHSU. During the initial visit, informed written consent or assent was obtained from all participants and their parents.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

People with autism often struggle with communication differences which traditional clinical measures and language tests cannot fully capture. Using language transcripts and audio recordings from 158 children ages 7 to 17, we showed that automated, objective language and voice measurements successfully predict the child's diagnosis. This methodology could help design improved outcome measures for research.

Keywords

autism; automated measures; communication; disfluency; natural language processing; pragmatic language; prosody; voice

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by a range of impairments in communication and social interaction, as well as restricted and repetitive patterns of interests and behavior (American Psychiatric Association, 2013). Recent reviews have placed the worldwide prevalence rate between 0.043% and 2.68% (Fombonne et al., 2021), with prevalence in the United States estimated to be 2.3% (Maenner et al., 2021). While reasons for the increasing prevalence rate are uncertain (Fombonne, 2018), this growing detection and awareness of autism requires thoughtful and respectful research to help facilitate communication between people with and without autism.

Communication differences are a core feature of autism with a highly variable presentation across individuals (Gernsbacher et al., 2016; Meir & Novogrodsky, 2020). Autistic children have variable expressive and receptive language skills and near-universal difficulties with pragmatic language (Andrés-Roqueta & Katsos, 2017; Tager-Flusberg & Caronna, 2007). Expressive and receptive language abilities are associated with cognitive level and several domains of executive functioning in autistic children (Akbar et al., 2013; Kjellmer et al., 2012). Ellis Weismer et al. (2018) showed that clinical evaluation of language fundamentals-4 (CELF-4; Semel et al., 2004) scores predicted the outcomes of nonverbal executive functioning tasks in autistic children. As well, Levinson et al. (2020) found that semantic, syntactic, and pragmatic language could predict parent- and teacher-reported social skills in autistic children. Autistic individuals also demonstrate a large variability in production of paralinguistic features, such as prosody (Globerson et al., 2015; Grossman et al., 2010); these differences are often only revealed using acoustic analysis (Loveall et al., 2021).

Current methods for quantifying language and voice characteristics are costly, difficult to administer, and of limited accuracy. Furthermore, the heterogeneity of language use among autistic people warrants development of accessible outcome measures that are easy to obtain, can accommodate a wide range of abilities, and are sensitive to change (Barokova & Tager-Flusberg, 2020). Conventional standardized tests like the CELF-4 and the peabody picture vocabulary test (PPVT; Dunn & Dunn, 2007) evaluate structural aspects of language and are useful as one source of information within a larger developmental framework (Tager-Flusberg et al., 2009); however, such clinical assessments must be augmented by outcome

measures derived from natural language samples and parent and individual reports. One such parent-reported measure is the commonly-used children's communication checklist-2 (CCC–2) (Bishop, 2003), yet there are limitations to the validity of such measures as illustrated in treatment studies that have shown caregiver reports to be susceptible to the placebo effect (Guastella et al., 2015). There are limited options for clinicians and caregivers to evaluate quantifiable vocal differences, as reflected by both the autism diagnostic observation schedule (ADOS) and social responsiveness scale (SRS) which include only one item each on voice quality (Asgari et al., 2021), despite the fact that prosodic differences are often cited as a specific characteristic of autism (Fusaroli et al., 2017; Paul et al., 2005). Some clinical measures are available to assess prosodic skills in children with ASD (Peppé et al., 2007; Shriberg et al., 2010); however, researchers have called for automated voice measures as a means of quicker, more accurate methodologies for classifying prosody in natural language (Bone et al., 2014).

Measurement of voice and language features has progressed with natural language processing (NLP) and voice analysis quantitative methodologies. Using a range of approaches, studies have examined linguistic features and voice characteristics both alone and in conjunction in order to predict diagnosis or screen for early signs of disease. For example, Alhanai et al. (2017) found that combining audio and text features increased detection of cognitive impairment in a large, longitudinal population study over both baseline demographic performance and over feature grouping. When applied to autism, NLP methods have also been used efficiently to characterize language profiles (Chojnicka & Wawer, 2020; Salem et al., 2021). Likewise, voice analyses have been performed showing discriminant prosodic patterns in autistic versus non-autistic participant groups (Asgari et al., 2021; Kiss et al., 2012; Li et al., 2019). The methodology of these initial investigations has been variable with respect to both the number and variety of features extracted from voice and language samples, and to the sampling contexts from which voice and language data were acquired. Some studies used small corpora and included a wide variety of both voice and language features to discriminate between autistic or typically developing children (Cho et al., 2019; Tanaka et al., 2014). In contrast, Parish-Morris et al. (2016) selected several specific features to analyze in a pilot study. They encouraged other researchers to add layers of analysis to a single, large, shared corpus of language transcripts, in order to facilitate a shared understanding of overarching patterns.

In previous work, we have used NLP methodologies and established their ability to differentiate between diagnostic groups (Adams et al., 2021; MacFarlane et al., 2017; Salem et al., 2021; van Santen et al., 2013). Salem et al. (2021) examined seven automated language measures (ALMs) and found them able to accurately classify already-diagnosed children into ASD and non-ASD groups even after controlling for age and IQ. Separately, we have also analyzed acoustic features of speech to evaluate their discriminant ability across diagnoses: Asgari, Chen, and Fombonne et al. (2021) found that a harmonic model of speech could accurately differentiate between children with ASD and those who were typically developing (TD); again, results held true after further adjustment on IQ.

In the present study, we aimed to confirm the performance of the automated voice and language measurements as previously established by Asgari et al. (2021) and Salem et

al. (2021), and further, to evaluate any potential gains derived from their combination. Specifically, our approach is to estimate the accuracy of a combined voice and language model in correctly classifying ASD status in a mixed sample of already-diagnosed children and a non-autistic comparison group, approaching this work with a streamlined, purposeful inclusion of established voice and language measures. This work (past and current) is not currently intended to be employed for diagnostic or screening purposes. Rather, our methodology uses already-established diagnosis as an external criterion against which we can validate (in the sense of discriminant validity) these voice and language measures. More than simply combining two types of measures, this approach offers an integrated, automated way of evaluating communication in general by accessing both voice and language characteristics.

Our study had three specific aims. First, we wanted to replicate the findings of the voice model described by Asgari et al. (2021) and of the language model described by Salem et al. (2021), and compare their relative discriminant validity using the same methodology within the same sample. Second, we wanted to measure where gains in accuracy could be obtained by combining all features from both the voice and language models when compared to that of single modality models. Third, we planned to conduct exploratory analyses of the clinical characteristics of participants who were or were not correctly classified by the automated models.

METHODS

Participants

Participants aged 7 to 17 years with either ASD, attention deficit hyperactivity disorder (ADHD), or TD were recruited for an fMRI study by community outreach and referrals from Oregon Health & Science University's specialty clinics. Parents completed a developmental and medical history survey at the first visit. Potential participants were screened for exclusion criteria: seizure disorder, cerebral palsy, pediatric stroke, history of chemotherapy, sensorimotor handicaps, closed head injury, thyroid disorder, schizophrenia, bipolar disorder, current major depressive episode, fetal alcohol syndrome, Tourette's disorder, severe vision impairments, Rett syndrome, current use of psychoactive medications, non-English speaker, or an IQ below 70. All children in the ASD and ADHD groups were directly assessed by experienced child psychiatrists and clinical psychologists who confirmed their diagnosis based on DSM-IV-TR criteria (American Psychiatric Association, 2000). The research diagnostic team reviewed results of the standardized diagnostic assessments (both videos and scored protocols) and used best estimate procedures to confirm the diagnosis. ASD was ruled out in the TD and ADHD groups based on the ADOS, clinical interview, and parent-completed autism questionnaires.

Of 289 screened participants, 104 were ultimately excluded from this study for not meeting strict diagnostic criteria and another 10 for failing to complete the initial assessment procedures, leaving a sample of 175 subjects (102 ASD, 45 ADHD, 28 TD) included in the main neuroimaging study. Of these, we excluded a further six subjects with ASD due to an untranscribable ADOS session, either for poor sound quality or non-compliance. Finally, for analysis, we excluded 11 more participants (eight ASD, three non-ASD): six

participants with undefined um proportion (for whom um and uh did not occur in the Friends, Relationships, and Marriage Conversation; see below), four participants who did not have IQ results, and one participant who had an acoustically-corrupted file. Sample characteristics for the final 158 participants (88 ASD, 70 non-ASD) are given in Table 1. More details about the study sample have been described previously by Salem et al. (2021), where it was initially used. A subset of this sample was also used in Asgari et al. (2021).

In this study, we will refer only to ASD and non-ASD groups. The ADHD and TD groups of the original study were combined into a non-ASD group for our analyses for two reasons: 1) preliminary analyses indicated that the TD and ADHD groups were not significantly different from each other for any of the ALMs; 2) our focus was on detecting autism-specific features and including a non-autistic clinical group alongside TD participants provided an opportunity to more adequately test for the specificity of the results with respect to the presence or absence of autism.

Instruments

The autism diagnostic observation schedule (ADOS) (Lord et al., 2000) is a semi-structured, standardized assessment in which a trained examiner engages participants in activities that are designed to elicit social and communication behaviors indicative of symptoms of ASD as defined in the DSM-IV-TR (American Psychiatric Association, 2000) and DSM 5 (American Psychiatric Association, 2013). All participants were administered the ADOS-2 Module 3. All ADOS interviews were administered by research assistants or a senior clinical psychologist trained to research reliability level. All administrations were videotaped and later transcribed. The microphone was suspended in a cage and placed on a table approximately 3 feet from where the child sat. This allowed the recording to accurately capture the child's language, even if, as is common in ADOS Module 3 administrations, the child sometimes moved around the room before being reminded to sit down. The social affect (SA) score (10 items; range 0–20) and the restricted and repetitive behavior (RRB) score (4 items; range 0–8) were used in these analyses (Gotham et al., 2009). Higher scores indicate more severe ASD symptoms.

Intellectual level of participants was estimated with a short form of the Wechsler Intelligence Scale for Children 4th Edition (WISC) (Wechsler, 2003). Three subtests were administered: Information, Block Design, and Vocabulary, allowing a full scale IQ to be estimated from the sum of scaled scores of the three subtests according to the formula set out by Sattler and Dumont (2004). Separate verbal and non-verbal IQ estimates cannot be derived from this short form.

Language characteristics and linguistic pragmatic abilities were assessed using the parent-completed children's communication checklist, second edition (CCC-2) (Bishop, 2003). The CCC-2 is a widely-used, 70-item standardized checklist of pragmatic and social communication behaviors applicable to children ages 4:0 to 16:11. Caregivers are asked to make a frequency judgment about how often behaviors occur on 4-point scale. The CCC-2 is divided into 10 subscales measuring: (A) speech, (B) syntax, (C) semantics, (D) coherence, (E) inappropriate initiation, (F) stereotyped language, (G) the use of context, (H) non-verbal communication, (I) social relationships, and (J) interests. The first four subscales (A–D)

evaluate articulation and phonology, language structure, vocabulary, and discourse; four other subscales (E–H) evaluate pragmatic aspects of communication as well as stereotyped language with atypical or unusual expressions and use of nonverbal communication like facial expressions, bodily movements, and gestures. The last two subscales (I and J) measure behaviors characteristic of children with ASD. Each subscale raw score is converted to age-standardized scores (mean = 10; SD = 3). A Structural Language scale score is derived by averaging scores A to D, and a pragmatic language scale score is obtained by averaging scores E to H. Lower scores are indicative of more problems.

Data

The steps to prepare and analyze the data are illustrated in Figure 1.

Transcription

All ADOS administrations were audio and video recorded. The audio was transcribed according to modified SALT guidelines (systematic analysis of language transcripts) (Miller & Iglesias, 2012) by a team of trained research assistants who were blind to the participants' diagnostic status and intellectual abilities. Speech was split into communication units (c-units) consisting of a main clause and any subordinate, modifying clauses, or of speech fragments that constitute a whole utterance such as responses to questions. Special attention was paid to notation of disfluencies (mazes) and any unintelligible speech was marked. Any disagreements between transcribers were resolved through discussion with a clinician and a consensus judgment. Transcribers participated in biannual consistency checks to review protocol and ensure continued standardization. Lab transcription guidelines are available upon request from the first author.

In their study examining voice measures, Asgari et al. (2021) found the Friends, Relationships, and Marriage Conversation task (hereafter "Friends task") of the ADOS to most strongly discriminate between ASD and TD groups. Likewise, Salem et al. (2021) also found the Friends task to provide a reliable source for language analysis. This conversational activity is administered in the second half of the ADOS, after the "warm up" period of the testing situation, and consistently yields the most c-units of all ADOS tasks. This activity has a high standardization of conversational questions, leading to good comparability between participants. The examiner uses pre-established, open-ended interview questions which are designed to facilitate the flow of conversation. Accordingly, we analyzed recordings of the Friends task of the ADOS for this study.

Audio pre-processing

The pipeline of harmonic feature extraction comprised three modules: audio segmentation, recording denoising, and harmonic feature extraction. First, the audio segmentation module split the recording into multiple frames (chunks of sound) based on timestamps. Only frames that belonged to the child and occurred during the Friends task were retained. Then, the denoising module removed background noise from the selected frames. Recordings from both groups contain similar, ubiquitous noise not caused by participants. If denoising is not performed, the acoustic features mainly represent the noise instead of participants and the classifier cannot distinguish between groups. Denoising was performed with a pre-trained,

fully automated, deep neural network-based denoiser (Defossez et al., 2020); thus, no parameters were identified or set for this step. Then, we extracted 10 harmonic features (described below) from each denoised frame. Finally, for each participant, for each harmonic measure, we averaged the extracted feature over all frames in the Friends task.

Automated language measures (ALMs)

We generated seven expressive language measures as described in Salem et al. (2021): Mean Length of Utterance in Morphemes (MLUM; calculated on all complete, fluent, and intelligible c-units), Number of Distinct Word Roots (NDWR; calculated on all complete, fluent, and intelligible c-units), um proportion (total number of *um* divided by the total number of *um* + *uh*), content maze proportion (number of content mazes [non-filler disfluencies] divided by the number of content mazes + the number of fillers), unintelligible proportion (number of partially or fully unintelligible c-units divided by the total number of c-units), c-units per minute (CPM; number of attempted c-units per minute), and repetition proportion (number of child words that are repeated in a set of two or more from the examiner's immediately preceding turn, divided by the total number of child words).

Automated voice measures (AVMs)

We analyzed 10 voice measures as described in Asgari et al. (2021): Cepstrum, delta cepstrum, delta delta cepstrum, log spectral entropy, F0, jitter, shimmer, harmonic-to-noise ratio (HNR), H1H2, root mean square (RMS). Cepstrum is the shape of the spectral envelope that is extracted from cepstral coefficients. Thirteen cepstral coefficients of each frame were augmented with their first- and second-order time derivatives. The derivatives are named as delta cepstrum and delta delta cepstrum. Spectral entropy is a proxy for cues related to voicing and quality. It can be used to characterize speechiness of the signal and has been widely employed to discriminate speech from noise. As such, we computed the entropy of the log power spectrum for each frame, where the log domain was chosen to mirror perception. F0 is the frequency of the vocal cords' vibration when voiced sounds are produced. It is also known as the fundamental frequency or "pitch." Jitter is the small cycle-to-cycle fluctuations in the glottal pitch period. Shimmer is the small cycle-to-cycle fluctuations in amplitude. Standard methods for calculating jitter and shimmer are sensitive to the pitch estimation. We therefore adopted a model-based approach for jitter and shimmer estimation, which produces measures of short-term pitch and amplitude variation for an estimation of jitter and shimmer, respectively (Asgari & Shafran, 2010). HNR is a quantity to measure the amount of noise in voice to assess the degree of hoarseness. H1H2 is an indicator of breathiness. RMS is the root means square energy of each frame and captures loudness.

Descriptive statistics for these seven ALMs and 10 AVMs are provided for each participant group, as well as between-group statistical comparisons, in the supplemental information (Table S1).

Statistical analyses

To evaluate how well the AVMs and ALMs each captured characteristics of ASD, we tested their ability to classify ASD status. We used a linear binary classifier called a support

vector machine (SVM). As noted by Ben-Hur and Weston (2010), SVMs are commonly used in machine learning and the sciences in general due to their high accuracy, ability to handle large sets of features, and flexibility to diverse types of data. They have been used successfully in previous research on speech and ASD, such as Pokorny et al.'s (2017) study on early identification of ASD using vocalizations and Asgari et al.'s (2021) study on acoustic features in ASD. SVMs find the optimal hyperplane to separate data into two classes, based on determining "support vectors," or critical data points that define the separating hyperplane. For the SVM training, we used a linear kernel and picked the best soft-margin constant C (an SVM hyperparameter) in each model: $C = 10$ for AVM, $C = 1$ for ALM and $C = 100$ for combined. We chose to report this SVM configuration due to its high performance, but comparisons of these results with non-linear SVMs and formal hyperparameter tuning can be found in Table S2. As is common practice for SVMs, we also normalized all our ALM and AVM features to be between 0 and 1 before they were passed to the SVM. First, we trained two separate SVMs to predict ASD status: one that used only the ALMs and a second that used only the AVMs. Then, to estimate the advantages of examining both voice and language features of children with ASD, we trained a third combined SVM to predict ASD status using the ALM and AVM measures together.

In order to limit over-fitting to the idiosyncratic characteristics of our participants, we performed leave-one-subject-out cross validation when evaluating SVM performance. This method involves repeated training of an SVM on all data except one participant, and then testing on the left out participant. When repeated for each subject, we can then use the combined predictions from each leave-out to estimate accuracy and other classifier evaluations. This form of cross validation is recommended by Saeb et al. (2017), due to its approximation of the clinically relevant use-case of diagnosis in unseen participants. We tested the robustness of this approach by comparing its results to those of a 10-fold cross-validation.

We calculated accuracy, sensitivity, and specificity of our SVM classification predictions. We also calculated receiver operating characteristic (ROC) curves for each model based upon the prediction probabilities from the three SVM models, and their corresponding area under the curve (AUC) values and 95% confidence intervals. To test whether the combined SVM model was significantly better performing than the individual AVM or ALM models, we employed the DeLong test for comparison of AUC from correlated ROC curves (DeLong et al., 1988). We further examined whether performance varied by sex and trained the combined model separately for boys and girls. Finally, in order to evaluate the respective contributions of each single feature comprised in the combined model, we used multiple support vector machine recursive feature elimination (mSVM-RFE; Duan et al., 2005).

To evaluate which participant characteristics the AVM and ALM features were picking up on, we used four clinical scores: two language scores (structural and pragmatic CCC-2 scales) and two measures of autism symptom severity (ADOS-2 SA and RRB scores). We also computed objective measures of language output as observed during the ADOS-2 administration of the Friends task. These Friends task indices were: total number of c-units, total number of words, and activity time (length in minutes); higher scores represent more language output. We then computed means for age, IQ, the four clinical scores, and

the three task indices of participants with ASD who were detected (correctly classified as ASD) or missed (incorrectly classified as non-ASD) for each of the three models. Statistical comparisons of mean scores were performed with the nonparametric Mann–Whitney/Wilcoxon rank sum test (Neuhäuser, 2011) for each model. Similar analyses were performed comparing detected and missed participants without ASD; these results are described in Table S3.

A p-value of <0.05 was retained as a level of statistical significance. All analyses were performed using R statistical computing software (RCoreTeam, 2017).

RESULTS

Discriminant validity of ALMs and AVMs

To address our first aim, we trained two separate SVMs to classify ASD status: one that used the AVMs and one that used the ALMs. The results for the language-only and voice-only models are summarized in Table 2. The voice SVM obtained an accuracy of 0.7215, with sensitivity 0.7500 and specificity 0.6857. This model thus performed satisfactorily at identifying ASD participants, but less so at identifying non-ASD participants. The ALM model obtained an accuracy of 0.7975, with sensitivity 0.7727 and specificity 0.8286. This model performed similarly well at correctly identifying participants with ASD and much better at correctly identifying non-ASD participants than the voice model. When evaluated with ROC analysis, the discriminant performance of the language model was significantly better than that of the voice model, as measured by the AUCs (see Table 2; DeLong p = 0.036).

Combining ALMs and AVMs

To address our second aim, we trained an SVM on both the AVMs and the ALMs to see how well the combination of these two types of features predicted diagnostic status. The classification results of the combined SVM are summarized in Table 2. This combined model obtained an accuracy of 0.8671, with sensitivity 0.8977 and specificity 0.8286, demonstrating substantial improvement over both the AVM SVM alone and the ALM SVM alone.

We calculated ROC curves for the voice model, the language model, and the combined model from the prediction probabilities. The three curves are plotted in Figure 2. AUC values and confidence intervals from these ROC curves are shown in Table 2. Again the combined model performed best, with an AUC of 0.9205 that was significantly higher than that of 0.7800 for the AVM model and that of 0.8748 for the ALM model (DeLong tests: $p < 0.001$ and $p = 0.033$, respectively). Using an alternative cross-validation method, the ROC AUC value for the 10-fold combined model was 0.9235 (95% CI 0.8801–0.967), in comparison to 0.9205 (95% CI 0.8768–0.9641) for the leave-one-out cross-validation, a non-significant difference ($p = 0.9219$).

We ran an SVM on only the boys in our sample ($n = 113$), and then only the girls ($n = 45$). The ROC AUC for the boys was 0.8901 (95% CI 0.8261–0.9541), which was not significantly different from the ROC AUC for the SVM with our full sample ($p =$

0.4428). The ROC AUC for the girls was 0.8067 (95% CI 0.6493–0.964), which was also not significantly different from the ROC AUC for the full sample ($p = 0.178$). Thus, performance of the combined model was not significantly affected by child sex.

For the combined model, we initially included all features—including those that did not significantly differentiate groups since they may still have value in a larger model. To find the optimal subset of features, we used mSVM-RFE on the combined model to rank the features across 10-fold cross validation and then find the combination with the lowest generalization error. We chose to use 10-fold cross validation for this step because the software could not be applied to leave-one-out cross validation. The optimal model used the following 8 features: content maze proportion, c-units per minute, RMS, F0, MLUM, Shimmer, um proportion, and Log Spectral Entropy. However, when comparing ROC AUC of the optimal set of features using 10-fold (0.9344, 95% CI 0.8931–0.9757) and the original 10-fold combined model (0.9235, 95% CI 0.8801–0.9670), they were not significantly different according to DeLong's test ($p = 0.1592$).

Clinical features associated with model classification performance

Age, IQ, clinical scores, and Friends task indices of participants with ASD who were detected or missed by each set of features (ALMs and AVMs) alone, and by the combined model, are summarized in Table 3.

For the AVM model, there were no statistical differences between ASD participants correctly or incorrectly classified. In the ALM model, compared to the missed participants, those who were detected had significantly lower language skills as measured by both the structural and pragmatic CCC-2 scores, and less language output during the Friends task as measured by the three task indices; they were also younger, on average, by one year. A comparison of scores of the participants correctly classified by the ALM and by the AVM models shows that those identified by the AVM model had consistently higher mean scores than their counterparts on the two CCC-2 scores and the three indices measuring the Friends task language output. Therefore, the voice-only model appeared to capture more verbally fluent participants than the language-only model. As a result, when voice was added to language in the combined model, correctly classified participants had higher average language scores and more language output compared to those correctly identified by the language features only.

Lastly, there remained nine subjects (10.2%) who were missed by the combined model. Because of this small number, statistical power was reduced for further between-group comparisons between detected and missed participants. Nevertheless, it is notable that of all subgroup means evaluated in Table 3, this subset of nine subjects had the highest means for IQ (105.1), number of c-units (108.9) and number of words (627.2). Both number of words and number of c-units showed significant differences between detected and missed participants. Taken together, these findings confirm a general trend for participants with milder phenotypes and more typical language to be those more easily missed.

While language indices were significantly associated with performance of several models, ADOS scores and IQ were never significantly associated with classification in any

model. This suggests that our methodology captures specific communication characteristics associated with autism that are unconfounded by cognitive level and autism severity.

Clinical and task indices associated with misclassification among participants without ASD showed that non-ASD subjects misclassified with the ALM model tended to be younger and have lower language skills and output, with few differences otherwise (Table S3).

DISCUSSION

In this study, we first replicated results from Salem et al. (2021) and Asgari et al. (2021) showing that a defined set of language- and acoustic-based measures generated from transcripts and audio recordings from an ADOS-2 task classified individuals with and without ASD better than chance alone. Compared to the voice model, the language model had better accuracy due to much improved specificity. Combining both sets of features in the predictive model resulted in significantly improved performance when compared to either single-modality model, shown by a higher level of sensitivity and improved AUC. In follow-up exploratory analyses, we examined the characteristics of the ASD participants who were missed or detected by each model, and found that the language model detected ASD individuals with more marked language and pragmatic difficulties, while the voice model detected ASD participants with higher language ability.

While both the voice and language models were better than chance at correctly classifying ASD and non-ASD children (as shown in Figure 2), the model using automated language measures performed better on its own than the voice model. However, these two models were picking up on children with different language patterns, as shown in Table 3. Compared to those detected by the language model, participants detected by the voice model had consistently higher levels of language skills as measured by both a parental report of their language over the last 6 months (the CCC-2) and by objective indices of language output during the ADOS-2 task administration. During the Friends task, they produced more words and c-units and stayed on the task for a longer period of time. Furthermore, in the voice model, the CCC-2 scores and the Friends task indices did not differentiate correctly or incorrectly classified participants, indicating that the voice model assesses paralinguistic features (e.g., prosody, loudness) that are truly independent of language level and output. This has important implications for evaluating the quality of communication skills in individuals with autism. In particular, it emphasizes the need to include voice analysis to detect communication differences among subjects with ASD who have achieved high language skills in their development. It also concurs with the clinical observation that voice atypicalities often persist among older individuals who have achieved the most “typical” outcomes and show otherwise few language anomalies.

The voice model’s low specificity (0.69) indicated that, with a prediction based on voice features only, too high a proportion of false positives would occur. Further work should investigate which, if any, of the several voice features included in our AVM model accounted for misclassifying individuals without ASD. For instance, it could be that loudness is a general non-specific feature of young children interviewed on sensitive topics (as those covered by the Friends task) in an unfamiliar setting; if so, removal of these non-

contributory features might improve the classification performance of voice-only models in the future. Likewise, the language model incorporated seven features that were shown in our previous work to have discriminant validity (Salem et al., 2021). However, they may have not equally contributed to the prediction of ASD since some measures were more discriminant than others (see Table S1). Indeed, results from our Recursive Feature Elimination analysis indicated that some features could possibly be excluded with small but insignificant gains in accuracy. Further investigations in more varied samples should be performed in the future to fully evaluate the predictive power of each single feature. We are in the process of generating and testing other features that could be added in future studies to potentially improve the accuracy of the language models.

The addition of voice to language features in the models led to a substantial gain of sensitivity, confirming that ASD individuals who were not initially detected by the language model were detected by the additional voice features. Notably, the specificity of the combined model was maintained at the same level as that of the language model, i.e. the addition of voice features did not increase misclassification among non-ASD children, despite that modality's tendency towards false positives. Overall, the combined model had excellent performance with only 10.2% of individuals with ASD being misclassified. Despite the reduced statistical power of comparisons involving this small subset, we still noted that while missed participants performed better on objective Friends task indices of language output, their parent-reported language ability (as described by the CCC-2) was not different from their correctly classified peers. There are several possible reasons for this discrepancy. First, the parents' CCC-2 scores could have been negatively biased in this subset due to their child's ASD diagnosis. Alternatively, it could be that despite their persisting language difficulties, these intellectually-able participants have developed compensation strategies that allowed them to display proficient language in the ADOS-2 testing situation. As well, as previously reported by Bone et al. (2012, 2014), examiners may have interacted with those participants in a way that facilitated and maximized their language output. Future research could elucidate this by including direct language testing and qualitative reviews of ADOS-2 administration.

The correctly classified ASD participants in the combined model had shorter transcripts, as defined by length of activity, number of words, and number of c-units. Despite the reduced language output, these participants still had sufficient language differences to be detected by our combined model. Therefore, it is unlikely that the missed participants were not detected due to a lack of language available for analysis or that extending the length of transcripts would reduce misclassification. On the other hand, one could argue that the high language output of the missed ASD participants in the combined model is a source of measurement error and that voice and language measures do not work well on longer language/voice samples. However, almost all of the voice and language measures were normalized by length of the sample through either being an average over frames or c-units, or a proportion that can be calculated on any length sample. The exception was NDWR, which could have a relationship with length of a sample; however, as shown in the supplemental information, NDWR was not an especially discriminant measure. Thus, it is highly unlikely that the measures did not perform well for high language output, and more likely that high IQ and older children had fewer communication difficulties, making them harder to detect.

Our findings are in line with existing literature (Cho et al., 2019; Parish-Morris et al., 2016; Tanaka et al., 2014) in showing that objectively measured voice and language features can correctly discriminate ASD and TD children. Our results outperformed those from previous studies. Tanaka et al. (2014) examined language features and speech (voice quality, pitch, intensity, rate) to predict ASD diagnosis and reported an accuracy of 68.8% in combining all features, with no evidence of differential performance by type of features (language: 70.3%; speech: 67.6%). However, the sample size was extremely small (4 ASD, 2 TD) making the findings unreliable. Our combined predictive model performed better than a combined language and speech model by Cho et al. (2019) in terms of both accuracy (0.867 vs. 0.757, respectively) and AUC in ROC analyses (0.920 vs. 0.754, respectively). Unfortunately, Cho et al. did not report individual voice or language model results. It ought to be noted that the sampling contexts used to generate language and voice data in those studies were different than ours. Tanaka et al. (2014) used recordings of parent–child semi-structured interactions; Cho et al. (2019) used unstructured conversations between children with ASD and young adult confederates. Examination of contextual effects should be vigorously pursued in future studies.

Studies have also varied in their exploration of factors associated with correct or incorrect overall classification. Cho et al. (2019) reported that adding IQ in their predictive model worsened its performance. In our study, a remarkable result was that autism severity and intellectual level were never significantly associated with misclassification in any of the three models we tested. If replicated, this result could confirm that our voice and language measurement methodology specifically taps aspects of communication skills that are not confounded by autism severity or developmental delay. If so, automated language and voice analyses could be deployed in samples with wide age, cognitive, and symptom distributions, at least when defined by the same range of clinical characteristics as those of our sample.

There are several strengths of our study. First, we evaluated voice and language features alone and in combination. This approach to communication analysis is more ecological and makes use of a full spoken sample as opposed to only one aspect (language features discarding the voice component, or vice versa). Second, compared to previous investigations, our sample size was large, encompassing an age range including school age and teenage years. Third, unlike traditional clinical research testing, the voice and language features in our methodology were generated blind to diagnostic status. Our methods are objective, quantitative, and user-friendly; insofar as transcripts and audio recordings are available, they can also be repeated many times without being vulnerable to learning/memory effects or attenuation.

We acknowledge several limitations as well. Our sample comprised children between the ages of 7 and 17; future language and voice studies should recruit adults and younger children. All participants had normal IQ range and all were fluent English speakers who could be assessed with the ADOS-2 Module 3. It is important for future studies to include individuals with lower IQ and less verbal participants. Participants were not tightly matched between groups which could result in reduced fit for analysis, although, as already noted, age, sex, and IQ were weakly related to model performance. The language and voice analyses were based on a limited set of robust, previously validated characteristics;

future analyses should expand on feature extraction through systematic testing of each component feature in order to improve the accuracy and sensitivity of each model. We do not know the full impact of denoising on voice features and results. Future studies should tightly control for microphone quality and placement, which may reduce the need for the denoising step. Our language and voice samples were derived from ADOS-2 administrations in a clinical research setting; therefore we cannot currently generalize these findings to different contexts. ADOS administration focuses on eliciting autistic features, thus it is possible that our results are overestimated compared to what would be obtained with more naturalistic conversational samples. It will be important to include other sampling contexts in future studies, aiming at minimizing interference by strangers and capturing naturally occurring conversations in the ecological niche of the participants. Although we are currently undertaking those studies, we do not yet know the reliability (test–retest) and short-term stability of the measures we generated from transcripts and audio recordings. Another limitation lies in the reliance of our language analysis methodology on manual transcription of recorded language samples, a costly and labor-intensive step. Moreover, though the transcribers went through extensive training and regular consistency reviews, we do not have transcription reliability results. However, we expect progresses in automatic speech recognition (ASR) methodology will bypass this inconvenient step and quickly increase the pace of research in this area. Future studies based on ASR transcription will need to re-evaluate performance of the models relying on ALMs.

Our findings provide preliminary but very encouraging results on the value of automated analysis of language and voice samples from children with ASD. In the future, our goal is to refine this methodology and translate it into the design of novel evaluation tools. These objective, automated, quantified, and user-friendly tools could be incorporated in randomized clinical trials or longitudinal studies to evaluate changes in communication skills as a function of intervention or maturation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award R01DC012033 (PI: Dr. E. Fombonne). The authors have no conflicts of interest to declare. The authors would like to thank Dr. Damien Fair’s neuroimaging team, which collected a portion of the baseline clinical data used in this study as part of their NIH funded studies R01MH115357 and R01MH086654. We thank, in particular, Michaela Cordova, Beth Calamé, Julia Painter, and Alicia Feryn, and we gratefully acknowledge the children and their families who participated in the studies.

DATA AVAILABILITY STATEMENT

Analysis code is available at https://github.com/alexandrasalem/combining_voice_language_asd_analysis. De-identified data is available upon request from the authors.

REFERENCES

Adams JR, Salem AC, MacFarlane H, Ingham R, Bedrick SD, Fombonne E, Dolata JK, Hill AP, & van Santen J (2021). A pseudo-value approach to analyze the semantic similarity of the speech of children with and without autism spectrum disorder. Frontiers in Psychology, 12, 668344. 10.3389/fpsyg.2021.668344 [PubMed: 34366986]

Akbar M, Loomis R, & Paul R (2013). The interplay of language on executive functions in children with ASD. Research in Autism Spectrum Disorders, 7(3), 494–501.

Alhanai T, Au R, & Glass J (2017). Spoken language biomarkers for detecting cognitive impairment. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 409–416). IEEE.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders: DSM-IV-TR. American Psychiatric Association.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders: DSM-V. American Psychiatric Association.

Andrés-Roqueta C, & Katsos N (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. Frontiers in Psychology, 8. 10.3389/fpsyg.2017.00996

Asgari M, Chen L, & Fombonne E (2021). Quantifying voice characteristics for detecting autism. Frontiers in Psychology, 12, 665096. 10.3389/fpsyg.2021.665096 [PubMed: 34557127]

Asgari M, & Shafran I (2010). Extracting cues from speech for predicting severity of Parkinson's disease. In 2010 IEEE international workshop on machine learning for signal processing (pp. 462–467). IEEE.

Barokova M, & Tager-Flusberg H (2020). Commentary: Measuring language change through natural language samples. Journal of Autism and Developmental Disorders, 50(7), 2287–2306. [PubMed: 29873016]

Ben-Hur A, & Weston J (2010). A user's guide to support vector machines. In Carugo O & Eisenhaber F (Eds.), Data mining techniques for the life sciences (Vol. 609, pp. 223–239). Humana Press.

Bishop DVM (2003). The Children's communication checklist: CCC-2. ASHA.

Bone D, Black MP, Lee CC, Williams ME, Levitt P, Lee S, & Narayanan S (2012). Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. Proceedings of Interspeech, 2012, 1043–1046.

Bone D, Lee C-C, Black MP, Williams ME, Lee S, Levitt P, & Narayanan S (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. Journal of Speech, Language, and Hearing Research, 57(4), 1162–1177.

Cho S, Liberman M, Ryant N, Cola M, Schultz RT, & Parish-Morris J (2019). Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. In Interspeech 2019 (pp. 2513–2517). ISCA.

Chojnicka I, & Wawer A (2020). Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. PLoS One, 15(3), e0229985. [PubMed: 32142537]

Defossez A, Synnaeve G, & Adi Y (2020). Real time speech enhancement in the waveform domain. arXiv:2006.12847 [cs, eess, stat].

DeLong ER, DeLong DM, & Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. [PubMed: 3203132]

Duan KB, Rajapakse JC, Wang H, & Azuaje F (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on Nanobioscience, 4(3), 228–234. [PubMed: 16220686]

Dunn LM, & Dunn DM (2007). PPVT-4: Peabody picture vocabulary test. Pearson Assessments.

Ellis Weismer S, Kaushanskaya M, Larson C, Mathée J, & Bolt D (2018). Executive function skills in school-age children with autism spectrum disorder: Association with language abilities. Journal of Speech, Language, and Hearing Research, 61(11), 2641–2658.

Fombonne E (2018). Editorial: The rising prevalence of autism. Journal of Child Psychology and Psychiatry, 59(7), 717–720. [PubMed: 29924395]

Fombonne E, MacFarlane H, & Salem AC (2021). Epidemiological surveys of ASD: Advances and remaining challenges. Journal of Autism and Developmental Disorders, 51, 4271–4290. [PubMed: 33864555]

Fusaroli R, Lambrechts A, Bang D, Bowler DM, & Gaigg SB (2017). "Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis": Vocal production in ASD. Autism Research, 10(3), 384–407. [PubMed: 27501063]

Gernsbacher MA, Morson EM, & Grace EJ (2016). Language development in autism. In Neurobiology of language (pp. 879–886). Elsevier.

Globerson E, Amir N, Kishon-Rabin L, & Golan O (2015). Prosody recognition in adults with high-functioning autism spectrum disorders: From psychoacoustics to cognition: Prosody recognition and psychoacoustics in ASD. Autism Research, 8(2), 153–163. [PubMed: 25428545]

Gotham K, Pickles A, & Lord C (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. Journal of Autism and Developmental Disorders, 39(5), 693–705. [PubMed: 19082876]

Grossman RB, Bemis RH, Plesa Skwerer D, & Tager-Flusberg H (2010). Lexical and affective prosody in children with high-functioning autism. Journal of Speech, Language, and Hearing Research, 53(3), 778–793.

Guastella AJ, Gray KM, Rinehart NJ, Alvares GA, Tonge BJ, Hickie IB, Keating CM, Cacciotti-Saija C, & Einfeld SL (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: A randomized controlled trial. Journal of Child Psychology and Psychiatry, 56(4), 444–452. [PubMed: 25087908]

Kiss G, van Santen JP, Prud'Hommeaux E, & Black LM (2012). Quantitative analysis of pitch in speech of children with neurodevelopmental disorders. In Thirteenth Annual Conference of the International Speech Communication Association (pp. 1343–1346). ICSA.

Kjellmer L, Hedvall Å, Fernell E, Gillberg C, & Norrelgen F (2012). Language and communication skills in preschool children with autism spectrum disorders: Contribution of cognition, severity of autism symptoms, and adaptive functioning to the variability. Research in Developmental Disabilities, 33(1), 172–180. [PubMed: 22093662]

Levinson S, Eisenhower A, Bush HH, Carter AS, & Blacher J (2020). Brief report: Predicting social skills from semantic, syntactic, and pragmatic language among young children with autism spectrum disorder. Journal of Autism and Developmental Disorders, 50(11), 4165–4175. [PubMed: 32215820]

Li M, Tang D, Zeng J, Zhou T, Zhu H, Chen B, & Zou X (2019). An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. Computer Speech & Language, 56, 80–94.

Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, Pickles A, & Rutter M (2000). The autism diagnostic observation schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 30(3), 205–223. [PubMed: 11055457]

Loveall SJ, Hawthorne K, & Gaines M (2021). A meta-analysis of prosody in autism, Williams syndrome, and Down syndrome. Journal of Communication Disorders, 89, 106055. 10.1016/j.jcomdis.2020.106055 [PubMed: 33285421]

MacFarlane H, Gorman K, Ingham R, Presmanes Hill A, Papadakis K, Kiss G, & van Santen J (2017). Quantitative analysis of disfluency in children with autism spectrum disorder or language impairment. PLoS One, 12(3), e0173936. [PubMed: 28296973]

Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, Furnier SM, Hallas L, Hall-Lande J, Hudson A, Hughes MM, Patrick M, Pierce K, Poynter JN, Salinas A, Shenouda J, Vehorn A, Warren Z, Constantino JN, … Cogswell ME (2021). Prevalence and characteristics of autism Spectrum disorder among children aged 8 years – Autism and developmental disabilities monitoring network, 11 sites, United States, 2018. Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002), 70(11), 1–16. 10.15585/mmwr.ss7011a1

Meir N, & Novogrodsky R (2020). Syntactic abilities and verbal memory in monolingual and bilingual children with high functioning autism (HFA). First Language, 40(4), 341–366.

Miller J, & Iglesias A (2012). SALT: Systematic analysis of language transcripts [Research version]. SALT Software.

Neuhäuser M (2011). Wilcoxon-Mann-Whitney test. In Lovric M (Ed.), International encyclopedia of statistical science (pp. 1656–1658). Springer.

Parish-Morris J, Liberman M, Ryant N, Cieri C, Bateman L, Ferguson E, & Schultz R (2016). Exploring autism spectrum disorders using HLT. In Proceedings of the third workshop on computational Lingusitics and clinical psychology (pp. 74–84). Association for Computational Linguistics.

Paul R, Shriberg LD, McSweeny J, Cicchetti D, Klin A, & Volkmar F (2005). Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. Journal of Autism and Developmental Disorders, 35(6), 861–869. [PubMed: 16283080]

Peppé S, McCann J, Gibbon F, O'Hare A, & Rutherford M (2007). Receptive and expressive prosodic ability in children with high-functioning autism. Journal of Speech, Language, and Hearing Research, 50(4), 1015–1028.

Pokorny FB, Schuller B, Marschik PB, Brueckner R, Nyström P, Cummins N, Bölte S, Einspieler C, & Falck-Ytter T (2017). Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach. In Interspeech 2017 (pp. 309–313). ISCA.

RCoreTeam. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Saeb S, Lonini L, Jayaraman A, Mohr DC, & Kording KP (2017). The need to approximate the use-case in clinical machine learning. GigaScience, 6(5), 1–9.

Salem AC, MacFarlane H, Adams JR, Lawley GO, Dolata JK, Bedrick S, & Fombonne E (2021). Evaluating atypical language in autism using automated language measures. Scientific Reports, 11(1), 10968. [PubMed: 34040042]

Sattler JM, & Dumont R (2004). Assessment of children: WISC-IV and WPPSI-III supplement. Jerome M. Sattler, Publisher Inc.

Semel EM, Wiig EH, & Secord W (2004). CELF 4: 4 screening test. Pearson, PsyhCorp.

Shriberg LD, Fourakis M, Hall SD, Karlsson HB, Lohmeier HL, McSweeny JL, Potter NL, Scheer-Cohen AR, Strand EA, Tilkens CM, & Wilson DL (2010). Extensions to the speech disorders classification system (SDCS). Clinical Linguistics & Phonetics, 24(10), 795–824. [PubMed: 20831378]

Sprent P (2011). Fisher exact test. In Lovric M (Ed.), International encyclopedia of statistical science. Springer. 10.1007/978-3-642-04898-2_253

Tager-Flusberg H, & Caronna E (2007). Language disorders: Autism and other pervasive developmental disorders. Pediatric Clinics of North America, 54(3), 469–481. [PubMed: 17543905]

Tager-Flusberg H, Rogers S, Cooper J, Landa R, Lord, Paul R, Rice M, Stoel-Gammon C, Wetherby A, and Yoder P (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. Journal of Speech, Language, and Hearing Research, 52(3):643–652.

Tanaka H, Sakti S, Neubig G, Toda T, & Nakamura S (2014). Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. In Proceedings of reality (pp. 88–96). Association for Computational Linguistics. The Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.

van Santen JPH, Sproat RW, & Hill AP (2013). Quantifying repetitive speech in autism spectrum disorders and language impairment. Autism Research, 6(5), 372–383. [PubMed: 23661504]

Wechsler D (2003). Wechsler intelligence scale for children–fourth edition (WISC-IV) (4th ed.). The Psychological Corporation.

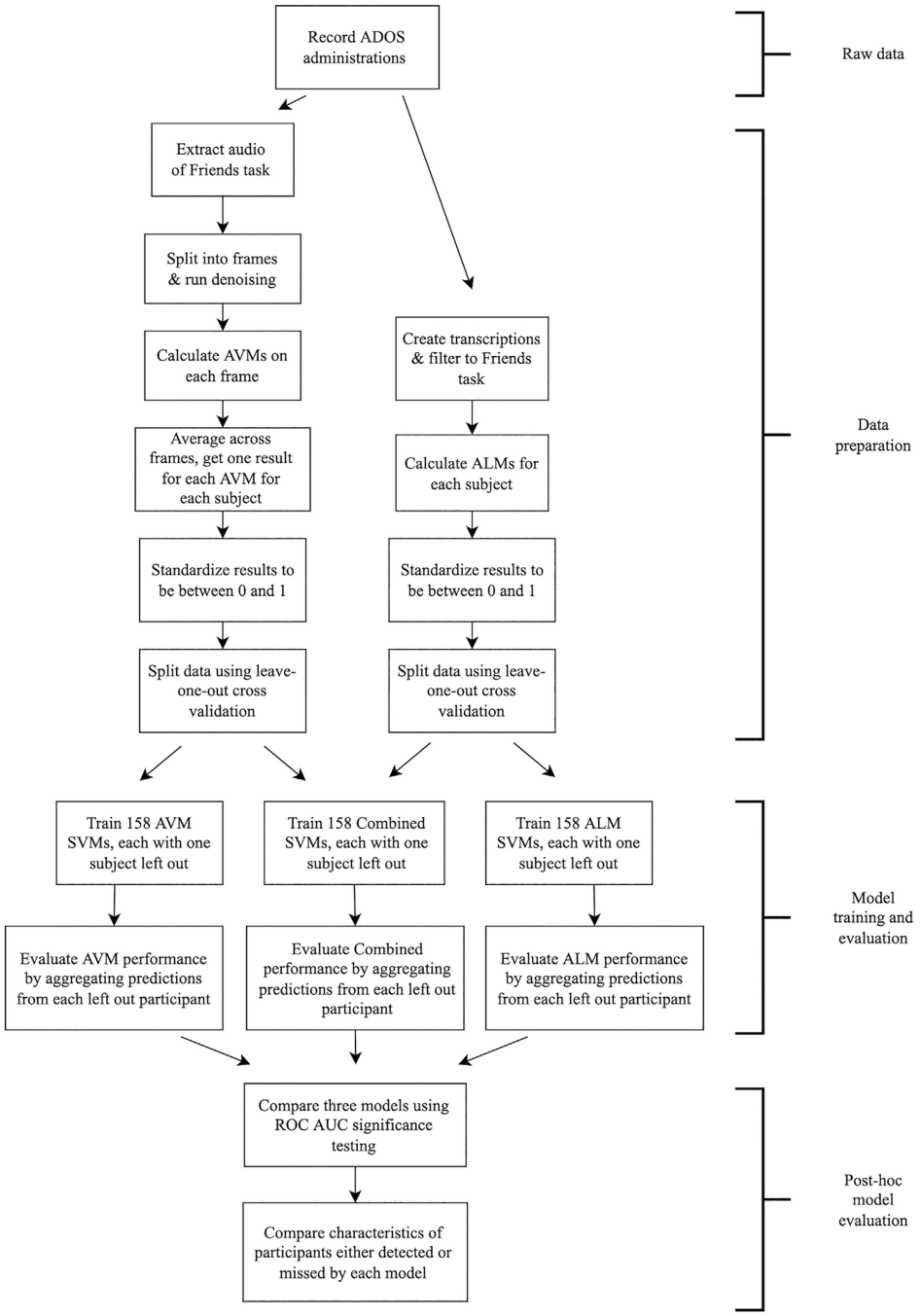

FIGURE 1.

Flowchart of study steps. ADOS, autism diagnostic observation schedule; ALM, automated language measure; AUC, area under the curve; AVM, automated voice measure; ROC, receiver operating characteristic; SVM, support vector machine

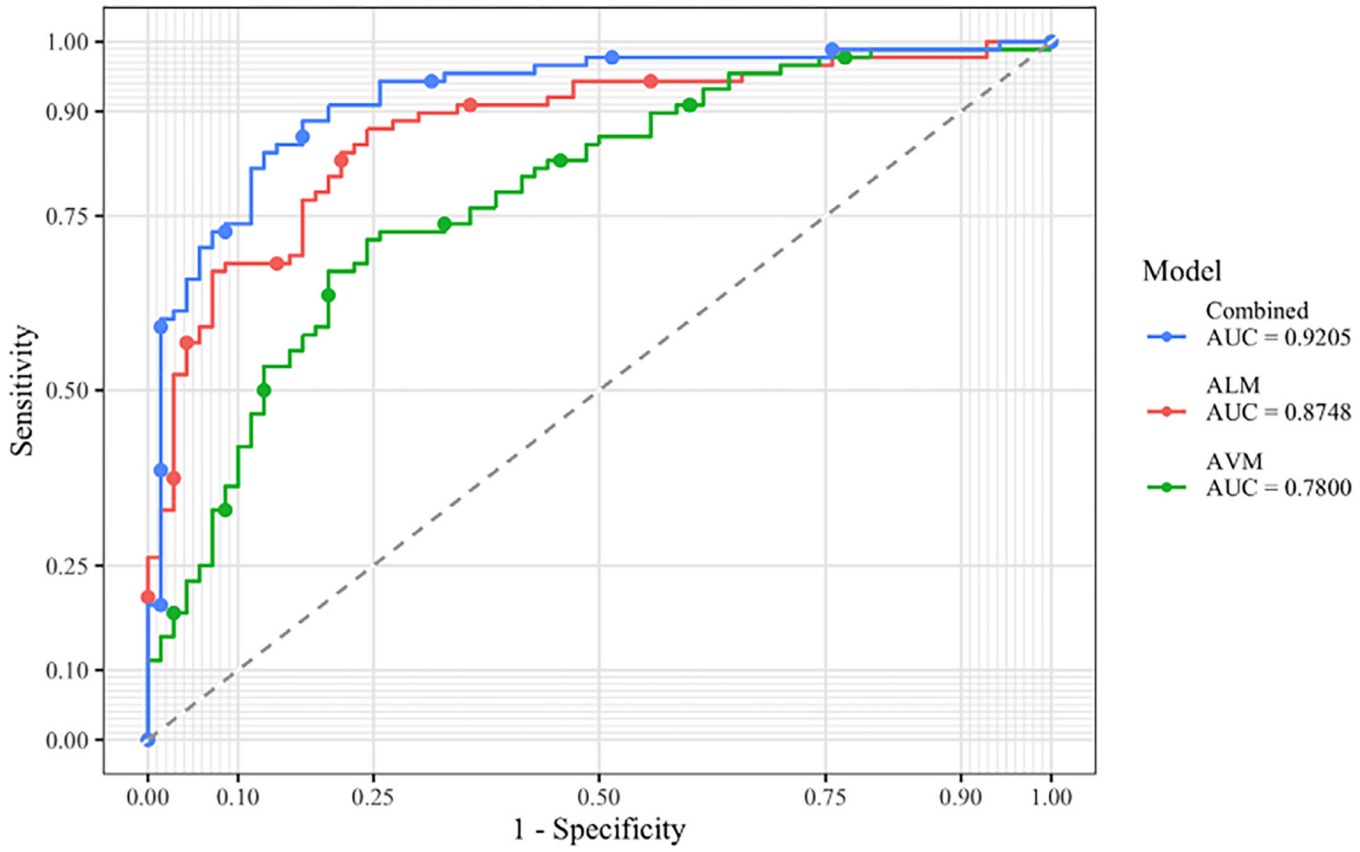

FIGURE 2.

ROC curves for SVM models, evaluated on class probabilities. ALM, automated language measure; AUC, area under the curve; AVM, automated voice measure

TABLE 1

Sample characteristics

	ASD	Non-ASD
	***n* = 88**	***n* = 70**
Sex, # male	73	40
Age in years (SD)	11.4 (2.1)	11.5 (1.7)
IQ (SD)	99.5 (19.9)	112.1 (12.9)
Ethnicity		
Hispanic, #	13	8
Non-Hispanic, #	74	62
Undeclared, #	1	0
Race		
American Indian/Alaska Native	3	0
Asian	0	2
Black	0	1
Native Hawaiian/Pacific Islander	1	0
White	70	59
More than one	11	7
Undeclared	3	1
ADOS-2 scores [a]		
RRB (SD)	3.48 (1.60)	0.46 (0.64)
SA (SD)	9.31 (3.53)	1.13 (1.58)
CCC-2 scores		
Structural score (SD)	6.53 (2.44)	10.21 (1.97)
Pragmatic score (SD)	4.92 (1.86)	10.24 (2.17)

Note: The Mann-Whitney/Wilcoxon rank sum test *p*-values for IQ, ADOS and CCC-2 scores were <0.001; the *p*-value for age was 0.6044. The Fisher's Exact Test for Count Data p-value for sex was 0.0006496 (Sprent, 2011).

Abbreviations: ADOS, autism diagnostic observation schedule; ASD, autism spectrum disorder; CCC, children's communication checklist; IQ, intelligent quotient; RRB, restricted and repetitive behavior; SA, social affect; SD, standard deviation.

[a]Three non-ASD participants were missing ADOS-2 scores.

TABLE 2

Model classification outcomes

	Accuracy	Sensitivity	Specificity	AUC (95% CI)
AVM	0.7215	0.7500	0.6857	0.7800 (0.7076–0.8525)
ALM	0.7975	0.7727	0.8286	0.8748 (0.8199–0.9297)
Combined	0.8671	0.8977	0.8286	**0.9205 (0.8768–0.9641)**

Note: Bold indicates the Combined SVM had significantly higher AUC than the AVM or ALM models alone.

Abbreviations: ALM, automated language measures; AUC, area under the curve; AVM, automated voice measures; CI, confidence interval.

TABLE 3

Averaged characteristics of detected and missed ASD participants in three predictive models

				Clinical scores				Friends task indices			
		Age	**IQ**	**CCC-2 Struc**	**CCC-2 Prag**	**ADOS RRB**	**ADOS SA**	**C-units**	**Words**	**Time**	***N***
AVM	Detected	11.19	98.36	6.56	5.09	3.53	9.70	75.27	378.50	6.60	66
	Missed	12.10	103.00	6.45	4.41	3.32	8.14	74.59	383.55	6.64	22
ALM	Detected	**11.11**	98.60	**6.09**	**4.72**	3.50	9.63	**67.53**	**323.28**	**6.24**	68
	Missed	**12.48**	102.65	**8.03**	**5.61**	3.40	8.20	**100.85**	**571.80**	**7.89**	20
Combined	Detected	11.29	98.89	6.50	4.86	3.48	9.46	**71.25**	**351.57**	6.48	79
	Missed	12.58	105.11	6.78	5.44	3.44	8.00	**108.89**	**627.22**	7.78	9

Note: Bold indicates the Mann–Whitney rank sum test *p*-value was less than 0.05. *N* is the count of ASD participants who were detected or missed by each model. C-units is the number of attempted c-units. Words is the total number of words. Time is length of activity, in minutes.

Abbreviations: ADOS, autism diagnostic observation schedule; ALM, automated language measures; AVM, automated voice measures; CCC, children's communication checklist; IQ, intelligence quotient; RRB, restricted and repetitive behavior; SA, social affect.