IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 30, 2020, accepted July 13, 2020, date of publication July 28, 2020, date of current version August 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012532

Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network

MARINA ENI[1], ILAN DINSTEIN[2,3,4], MICHAL ILAN[3,4,5], IDAN MENASHE[4,6], GAL MEIRI[4,5], AND YANIV ZIGEL[1], (Senior Member, IEEE)

[1]Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beersheba 8410501, Israel
[2]Department of Brain and Cognitive Sciences, Ben-Gurion University of the Negev, Beersheba 8410501, Israel
[3]Department of Psychology, Ben-Gurion University of the Negev, Beersheba 8410501, Israel
[4]National Autism Research Center of Israel, Ben-Gurion University of the Negev, Beersheba 8410501, Israel
[5]Pre-School Psychiatry Unit, Soroka University Medical Center, Beersheba 8457108, Israel
[6]Public Health Department, Ben-Gurion University of the Negev, Beersheba 8410501, Israel

Corresponding author: Marina Eni (marinamu@post.bgu.ac.il)

This work was supported in part by the Israel Science Foundation under Grant 961/14 and by the BGU Research Authority.

ABSTRACT Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that involves difficulties in social communication. Previous research has demonstrated that these difficulties are apparent in the way ASD children speak, indicating that it may be possible to estimate ASD severity using quantitative features of speech. Here, we extracted a variety of prosodic, acoustic, and conversational features from speech recordings of Hebrew speaking children who completed an Autism Diagnostic Observation Schedule (ADOS) assessment. Sixty features were extracted from the recordings of 72 children and 21 of the features were significantly correlated with the children's ADOS scores. Positive correlations were found with pitch variability and Zero Crossing Rate (ZCR), while negative correlations were found with the speed and number of vocal responses to the clinician, and the overall number of vocalizations. Using these features, we built several Deep Neural Network (DNN) algorithms to estimate ADOS scores and compared their performance with Linear Regression and Support Vector Regression (SVR) models. We found that a Convolutional Neural Network (CNN) yielded the best results. This algorithm predicted ADOS scores with a mean RMSE of 4.65 and a mean correlation of 0.72 with the true ADOS scores when trained and tested on different sub-samples of the available data. Automated algorithms with the ability to predict ASD severity in a reliable and sensitive manner have the potential of revolutionizing early ASD identification, quantification of symptom severity, and assessment of treatment efficacy.

INDEX TERMS Audio signals, autism, autism diagnostic observation schedule, autism spectrum disorder, convolutional neural network, deep neural network, early detection, outcome measure, pitch, speech, symptom severity, treatment efficacy, zero crossing rate.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder that is diagnosed by the presence of social communication impairments, repetitive behaviors, and confined interests [1]. The vast majority of ASD children exhibit speech and expressive language abnormalities, which range from a total lack of speech (i.e., non-verbal children) to those who develop normal vocabulary and syntax, but exhibit difficulties with the use of appropriate prosody and pragmatics [2]. This heterogeneity is apparent in the variable scores that children with ASD receive in standardized language assessments [1].

Early studies that have examined speech in children with ASD were based on small samples and manual analysis of short speech recordings. These studies have revealed that a considerable number of children with ASD exhibit expressive language delays often involving a prolonged pre-verbal stage [3]. Of those who do develop speech, many exhibit echolalia

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

(i.e., repeating words or phrases for no apparent reason) [4], including children with ASD who are deaf and use sign language [5]. Abnormal prosody is also very common [6] including reports of increased pitch variability and pitch range [7], slower speech rate [8], and prolonged word production [9].

Using automated speech-processing techniques to identify and quantify these and other speech abnormalities would be of great clinical utility. For example, such techniques could be used to assess early ASD risk, measure the severity of symptoms, and quantify their improvement or deterioration over time (e.g., in response to treatment). Furthermore, early intervention can successfully target expressive language capabilities in children with ASD, with the goal of improving general outcome [10].

With this in mind, several recent studies have utilized automated speech processing algorithms to examine speech recordings of English-speaking children with ASD. Most prominent are those that have used the Language Environment Analysis (LENA) system [11], a commercially available system that enables long audio recordings of the children in their natural surroundings for several days. These recordings are automatically segmented into speech segments that are grouped by speaker, and the child is identified by his relatively high pitch. Studies utilizing LENA have reported that children with ASD differ significantly from typically developing (TD) children in the degree of voicing [11], amount of vocalizations [12], length of syllables [11], clarity of formant transitions [11], and amount of vocal reciprocity (i.e., conversational turn-taking) [13].

While these findings are encouraging, LENA is not an open-source software and it is, therefore, not possible to further develop the LENA algorithms nor assess their utility in estimating ASD severity. One recent study used a Deep Neural Network (DNN) algorithm to estimate ASD severity from Autism Diagnostic Observation Schedule (ADOS) recordings of 33 English-speaking children [14]. They estimated the Social Affect (SA) scores of the ADOS, which specifically assesses the children's social abilities, and successfully explained ∼40% of the variability in these scores (i.e., $R^2 = 0.4$). These results suggest that it may be possible to predict ASD severity from relatively limited recordings of speech performed during the 1-hour ADOS assessment.

Here, we compared the ability of six different algorithms (two regression and four DNN models) to estimate ADOS scores from recordings of ADOS assessments performed with Hebrew speaking children. We extracted a variety of prosodic and conversational speech features from each recording and used them to train and test each of the algorithms, using a balanced cross validation approach.

II. EXPERIMENTAL SETUP

We selected recordings of 72 ADOS sessions that were performed at the National Autism Research Center of Israel (www.autismisrael.org), a collaborative project between Ben-Gurion University of the Negev (BGU), Soroka University Medical Center (SUMC), and other universities and medical centers in Israel. The recordings were performed with one microphone (CHM99, AKG, Vienna), which was located 1-2 m from the child (Fig. 1) and connected to a sound card (US-16 × 08, TASCAM, California). Each ADOS session lasted ∼40-minutes (41.6 ± 12.4 min) and was recorded at a sampling rate of 44.1 kHz (down sampled to 16 kHz). Of the 72 children included in the study, 56 had a diagnosis of ASD, 10 were referred with a suspicion of ASD, but received other diagnoses (e.g., language or developmental delays), and 6 were typically developing controls (Table 1).

FIGURE 1. The ADOS recording room at the National Autism Research Center.

All children completed a full clinical assessment according to DSM-5 [15] criteria as well as an ADOS (second edition) assessment using the toddler's module (n=10), module 1 (n=19), module 2 (n=31), or module 3 (n=12). The selection of the module depends on the age and language capacity of the child. The ADOS is a semi-structured assessment where a clinician administers specific tasks, observes the behavior of the child, and scores their behavior. The total ADOS score is in the range of 0-30 with higher scores indicating more severe symptoms. The total ADOS score is composed of SA (0-22) and Restricted and Repetitive Behavior (RRB, 0-8) scores, which can be standardized into comparison scores that enable comparison of ADOS scores across multiple ages and ADOS modules [16].

TABLE 1. Children characteristics.

	ASD (n=56)	Typical development (n=6)	Other diagnoses (n=10)
Age (months)	50.3±14.8	38.3±14.9	56.2±15.9
#of boys	49	5	9
ADOS score	14.1±5.7	1±1.3	7.5±3.9
SA score	10.1±4.9	0.3±0.8	5.9±3.4
RRB score	4.1±1.7	0.7±1.2	1.6±1.1
Cognitive score	77.7±17.7	116.3±8.4	87.7±9.3

III. METHODS

A. MANUAL LABELING OF SPEECH INTERVALS

We developed in-house software with a Graphical User Interface (GUI) and performed manual labeling of audio intervals containing speech (and/or other sounds, such as: crying, yelling, and mumbling) of the child, therapist 1, therapist

2 (some sessions had two clinicians), parent, simultaneous speech, and noise (e.g., chair being moved). All remaining intervals were automatically labeled as silence (i.e., background noise). Simultaneous speech was defined as speech of more than one speaker at a time.

B. AUTOMATIC DETECTION OF VOCAL SEGMENTS

Intervals of speech often contained multiple vocal segments (e.g., multiple utterances) separated by silence. To more accurately isolate vocal segments of individual speakers, we performed the following steps. First, we removed the dc of each audio recording (i.e., entire session). Second, we divided the audio signal into 40 ms frames with 30 ms overlap (i.e., frame rate of 10 ms), and computed the energy (in dB) in each frame (1):

$$E(i) = 10 \cdot \log_{10}\left(\frac{1}{N}\sum_{j=1}^{N} x(j)^2\right) \quad (1)$$

where i is the frame index, j represents the sample index in the i^{th} frame, and N is the total number of samples in a frame.

Third, we defined a baseline energy level, E_b, (in dB), as the most frequent energy level (i.e., background noise) within the audio interval and its vicinity ($\pm$ 20 s). Fourth, we defined the start of each vocal segment, Seg_{start}, as the frame where the energy level was 90% above E_b ($th_1 = 10 \cdot \log_{10}(1.9) + E_b$) for at least 50 ms (see Fig. 2). Fifth, we defined the end of the vocal segment, Seg_{end}, as the frame where the energy level was 10% above E_b ($th_2 = 10 \cdot \log_{10}(1.1) + E_b$) for 50 ms. Segments that were shorter than 110 ms (too short to contain an utterance) or longer than 3 s (too long to contain a well-formed phrase) were excluded from further analysis as also performed by the LENA's algorithm [11]. We found that <1.5% of vocal segments were excluded by these criteria. These steps and criteria allowed us to isolate individual vocal segments within each audio interval.

C. FEATURE EXTRACTION

We extracted 60 features from the vocal segments of each recording using custom written code and the PRAAT software for analysis of pitch and formants [17].

1) PROSODIC AND ACOUSTIC FEATURES

Pitch (F_0): The fundamental frequency generated by the child's vocal folds [18]. Pitch was calculated for each frame (40 ms length, 10 ms frame rate). Frames with pitch values below the voicing threshold were excluded from further analysis (voicing threshold was set to 0.45, which is a default value for normałlaryngeal speech [19]). The following 12 features were calculated for each child/recording:

1. Mean pitch across frame of all vocal segments.
2. Variance of pitch across frames of all vocal segments.
3. Variance of pitch divided by the mean pitch of the segment (i.e., pitch coefficient of variation). Mean was computed across vocal segments.

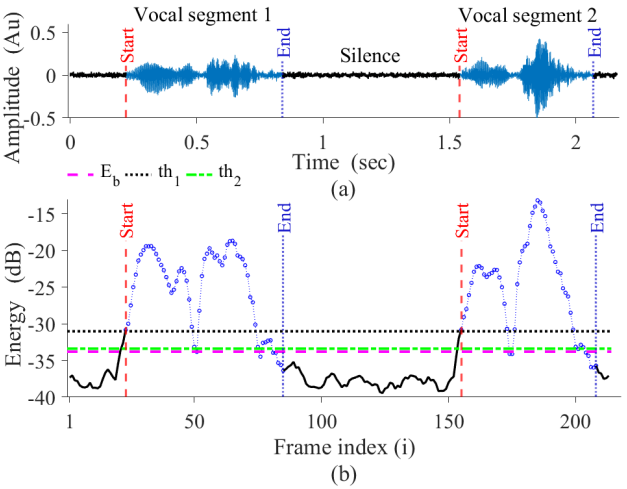

FIGURE 2. Example of a child's speech interval, which was segmented into two vocal segments. (a) The speech interval in the time domain. (b) The energy of the interval. Vertical lines in both panels mark the start (red dashed line) and end of individual vocal segments (blue dotted line). Horizontal lines mark the baseline energy level (dashed magenta line), and the two energy thresholds: th_1 (dotted gray line) and th_2 (dashed-dotted green line).

4. Mean minimum value of pitch across vocal segments.
5. Variance of minimum pitch across vocal segments.
6. Mean maximum value of pitch across vocal segments.
7. Variance of maximum pitch across vocal segments.
8. Mean pitch across voiced segments only. Voiced segments were defined as those where >60% of frames had pitch in the range of 60-1600 Hz [11].
9. Variance of pitch across voiced segments only.
10. Variance of mean pitch across voiced segments only.
11. Mean autocorrelation value from frames with pitch across all vocal segments.
12. Variance of autocorrelation value across frames with pitch.

Formants: The resonant frequencies, or formants, are mainly determined by the size and shape of the vocal tract, including the tongue, pharynx, and laryngeal, oral and nasal cavities [20]. Audible formant transitions occur when the vocal tract moves from a consonant closure to a vowel or vice versa [21]. The first two formants, and their bandwidths, were calculated for each frame, and the following features were extracted:

13. Mean first formant across frame of all vocal segments.
14. Variance of first formant across frames of vocal segments.
15. Mean second formant across vocal segments.
16. Variance of second formant across vocal segments.
17. Mean absolute difference between the two formants across frames of all vocal segments.
18. Variance of absolute difference between the two formants across frames of all vocal segments.
19. Mean bandwidth of 1$^{\text{st}}$ formant across vocal segments.
20. Variance of 1$^{\text{st}}$ formant bandwidth across vocal segments.

21. Mean 2nd formant bandwidth across vocal segments.
22. Variance of 2nd formant bandwidth across vocal segments.

Spectral slope: We fit a liner function to the FFT magnitude of two frequency ranges: 20-500 Hz and 500-1500 Hz [22], [23] for each frame in each vocal segment. These two frequency ranges were selected, because it was found that they contain information regarding voice quality and emotional states [22], [24]. This resulted in two spectral slopes per frame. The following features were then extracted:

23. Mean slope in 20-500 Hz across frame of vocal segments.
24. Variance of slope in 20-500 Hz across vocal segments.
25. Mean slope in 500-1500 Hz across frames of vocal segments.
26. Variance of slope in 500-1500 Hz across vocal segments.
27. Mean slope in 20-500 Hz across voiced segments only.
28. Mean slope in 20-500 Hz across unvoiced segments only.
29. Mean slope in 500-1500 Hz across voiced segments only.
30. Mean slope in 500-1500 Hz across unvoiced segments only.

Jitter: Is a measure of the cycle-to-cycle variations of fundamental frequency, which is commonly used for speaker identification and voice pathology [25]. We calculated jitter for frames with a defined pitch (F_0 was not zero).

31. Mean jitter across frame of all vocal segments.
32. Variance of jitter across frames of all vocal segments.

Energy: Since speech energy (i.e., volume) can be influenced from the distance between the child and the microphone, we normalized the energy (i.e., all frames) by the maximum energy (i.e., frame with maximal energy) of each recording (E_{norm}). The delta energy (ΔE_{norm}) and the delta-delta energy ($\Delta\Delta E_{norm}$) were calculated as well yielding the following eight energy features:

33. Mean of E_{norm} across vocal segments.
34. Mean change in energy across consecutive frames (i.e., first derivative, ΔE_{norm}), across vocal segments.
35. Mean second derivate ($\Delta\Delta E_{norm}$) across vocal segments.
36. Mean absolute value of ΔE_{norm} across vocal segments.
37. Variance of E_{norm} across vocal segments.
38. Variance of ΔE_{norm} across vocal segments.
39. Variance of $\Delta\Delta E_{norm}$ across vocal segments.
40. Variance of absolute ΔE_{norm} across vocal segments.

ZCR: We quantified the Zero Crossing Rate (ZCR) in each frame [26] in each vocal segment and extracted the following features:

41. Mean ZCR across vocal segments.
42. Variance of ZCR across vocal segments.
43. Mean ZCR across voiced segments only.
44. Variance of ZCR across voiced segments only.
45. Mean ZCR across unvoiced segments only.
46. Variance of ZCR across unvoiced segments only.

2) CONVERSATIONAL FEATURES

Vocalization rate: We counted the number of child vocal segments per minute in each recording.

47. Mean vocalization rate.
48. Variance of vocalization rate.

Duration: We computed the duration/length of each of the vocal segments. The following features were extracted:

49. Mean length of vocal segments.
50. Variance in the length of vocal segments.
51. Mean length of voiced vocal segments.
52. Variance in the length of voiced vocal segments.
53. Mean length of unvoiced vocal segments.
54. Variance in the length of unvoiced vocal segments.
55. Ratio between the mean length of voiced and unvoiced segments.

Turn-taking: We defined turn-taking as cases where a therapist's vocal segment was followed, within 2 s, by a child's vocal segment. This yielded the following features:

56. Mean response time of the child (i.e., time from end of the therapist's segment).
57. Variance of the response time.
58. Mean number of turn-takings in one minute (conversational rate).
59. Variance of the conversational rate.

#Segments: The amount of child vocalizations.

60. Total number of the child's vocal segments.

D. MODELS FOR PREDICTING ADOS SCORES

We built six models for predicting individual ADOS scores:

1) MULTPLE LINEAR REGRESSION

We divided the training dataset into 5 random groups (folds), trained the multiple linear regression model using 4 of the groups, and calculated the Root Mean Square Error (RMSE) between the true ADOS scores and the predicted scores in the left out group. We performed this procedure 5 times (i.e., 5-folds cross validation) and used a Sequential Forward Feature Selection (SFS) strategy [27] to rank the features according to their ability to reduce the RMSE between the true ADOS scores and the predicted scores. Hence, each iteration resulted with a ranking of the 60 features from best to worst. We then computed the RMSE as a function of the number of features included (for each of the 5 iterations) and calculated the mean RMSE across iterations. We found that, on average, $M^* = 15$ features yielded the lowest mean RMSE (Fig. 3). We then trained a final model using the 15 most popular features across the 5 iterations/folds. We utilized the entire training dataset (70% of the original data) and tested the ability of the model to predict ADOS scores from the independent testing dataset (30% of the original data).

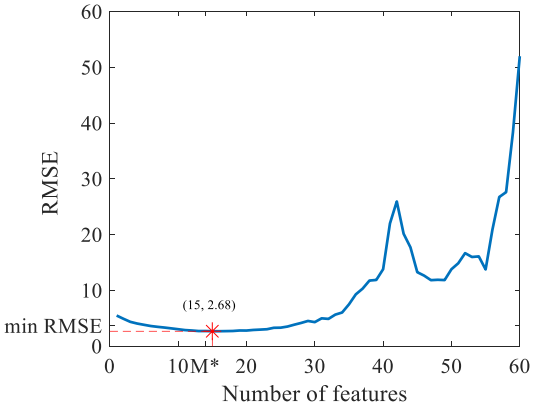

FIGURE 3. Feature selection. Demonstration of the average RMSE value when including an increasing number of features in the model. The optimal number of features was 15 and is marked with the red star, achieving a minimum average RMSE of 2.68 in the training dataset.

2) SUPPORT VECTOR REGRESSION (SVR)

SVR is another commonly used regression model for analyzing speech [28]–[30]. We estimated ADOS scores using SVR while applying either a linear, Gaussian, polynomial, or radial basis function kernel, and z-normalizing each of the features. The results of the SVR model with the linear kernel were superior to the other kernels; hence, we report only the results from the linear kernel.

3) FULLY CONNECTED DEEP NEURAL NETWORKS

We built three Fully Connected (FC) DNN models that differed in their input arrangement (feature arrangement, see below) and the structure of their input layer. The three FC models contained one input layer, 3 FC hidden layers, and a final output FC layer that yielded the ADOS predictions for each of the recordings/children (Fig. 4). A dropout of 0.5 was set between each two FC layers, in order to minimize the risk of overfitting [31].

Model 1 - Feature Vector for Each Session: The input for this model was a single vector for each recording session containing the values of the 60 features described above (Table 2) as computed across all vocal segments (Fig. 4a). The model was trained for 2000 epochs using batches of 8 samples in each training iteration.

Model 2 - Feature Vector for a Combination of Vocal Segments: DNN algorithms benefit from the availability of more data samples. To increase the number of data samples we selected sub-groups of 10 sequential vocal segments (of the child) and computed a subset of the features (1-46 and 49-50) for these sub-groups. We performed this procedure 100 times, selecting random sub-groups of sequential segments from each recording. This increased the number of available training samples from 51 (i.e., one per child in the training dataset) to 5100 (i.e., 100 per child) and enabled us to train the model with batches of 64 samples. The input to the DNN model was a vector containing the values of 48 features that were computed for each sub-group of 10 segments (Fig. 4b). The DNN was trained, for 4000 epochs, with individual vectors/samples such that each sample was associated with the ADOS score of the relevant child. The 48 selected features excluded the conversational features and included only features pertaining to the child's speech characteristics. The final output layer of the DNN yielded the predicted ADOS score of each sample. To generate the predicted ADOS score per child we computed the mean ADOS score across the samples of each child.

Model 3 - Feature Matrix for Combinations of Vocal Segments: Here we applied the same logic as in Model 2 but combined the 100 selected samples into a single input matrix. This yielded an input matrix of 100 $\times$ 48, one matrix per child. This model was trained with 51 samples/matrices in batches of 8 for 4000 epochs. Note, however, that each sample was a matrix that contained information from 100 randomly selected sub-groups of 10 sequential vocal segments (i.e., increasing the amount of information available in each sample). The final output layer of the DNN yielded the predicted ADOS score of the session/child (Fig. 4c).

4) CONVOLUTIONAL NEURAL NETWORK (CNN)

We also built a CNN model, which has previously been shown to accurately identify different aspects of speech intonation and prosody [32]–[35]. As in DNN model 3, here we used an input feature matrix with a size of 100 $\times$ 48 for each child. The model composed of two one-dimensional convolutional layers (Fig. 4d), with 256 filters (f) and a kernel size of 3 (k). A one-dimensional max pooling layer with pooling size of 3 (p) was evaluated between the two convolutional layers in order to help remove variability in the time-frequency domain that exists due to speech variability within each recording [36]. Next, four FC hidden layers with reducing dimensionality (1024-512-256-128 units) were used and followed by the output layer. We applied a ReLu activation function to the output of each hidden FC layers and the two convolutional layers, and a dropout of 0.5 after the first two FC layers. This model was trained using batches of 4 samples for 400 epochs.

The final number of FC layers, convolutional layers, and number of units in each architecture were chosen by tuning the models with different combinations of parameters. In addition, we selected optimal batch size and learning rate parameters for each of the DNN models by testing all pairs of the following combinations: batch size $\in$ {4,8,16} for models 1,3 and CNN, and {32, 64} for model 2; learning rate $\in$ {1e-5, 5e-5, 1e-4, 5e-4}. The final selected parameters for each model are shown in Table 3.

Since our algorithm was developed to solve a regression problem, all four models were trained using the MSE loss function, and RMSProp (Root Mean Square Propagation) optimizer [37].

In the training step, all DNN and CNN models received a feature vector/matrix from children in the training dataset and a target vector of their true ADOS scores. The scores were normalized by 30, to derive a target vector with a range of [0, 1]. After training was completed, the test dataset was evaluated, yielding the predicted ADOS scores (these were multiplied by 30 for comparison with the true ADOS scores).

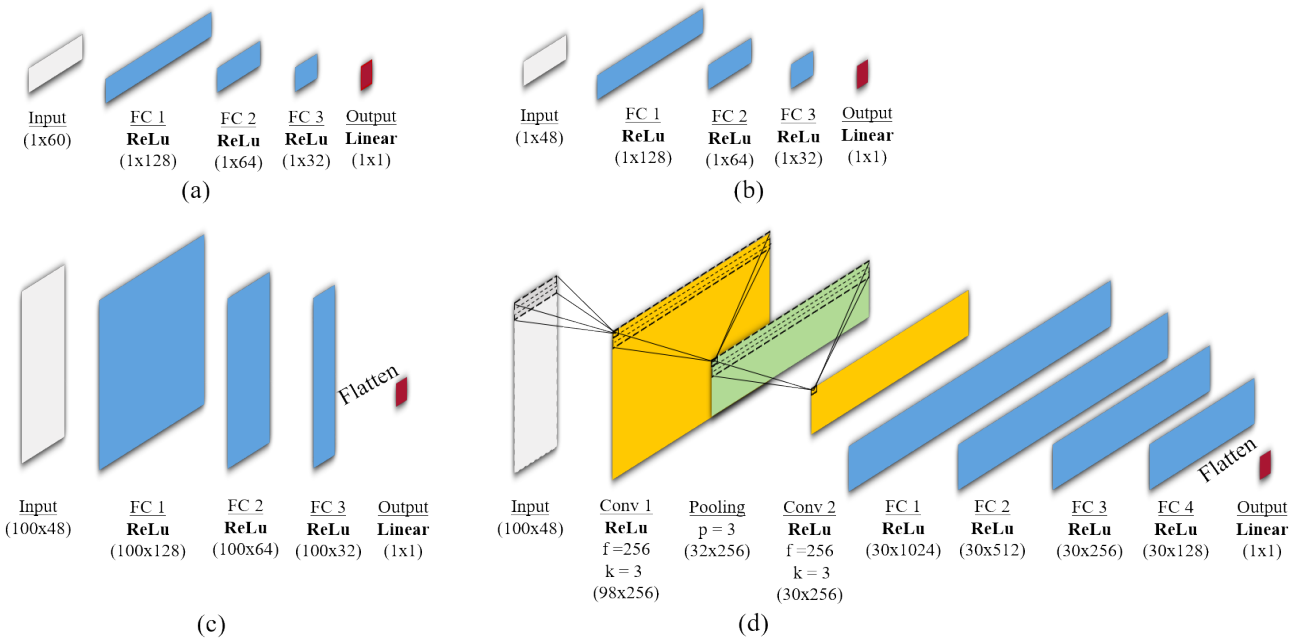

FIGURE 4. DNN architectures for ADOS estimation. (a) FC-DNN model 1 – a model with a single 60-dimensional feature vector as input. (b) FC-DNN model 2 – a model with a single 48-dimensional feature vector as input. (c) FC-DNN model 3 – a model with a 100 × 48-dimensional feature matrix as input. (d) CNN model – convolutional model with a 100 × 48-feature matrix as input, with 256 filters and kernel size of 3 in the convolutional layers, and a pooling size of 3.

TABLE 2. Features used in each of the DNN models.

	FC Model 1	FC Model 2	FC Model 3	CNN Model
Features included	1-60	1-46, 49-50	1-46, 49-50	1-46, 49-50
Total	60	48	48	48

TABLE 3. Hyper parameters of the DNN models.

	FC Model 1	FC Model 2	FC Model 3	CNN Model
Learning rate	5e-5	5e-6	1e-5	1e-5
Batch size	8	64	8	4
# Parameters	18,817	16,641	19,809	1,190,017

E. DATA ANALYSIS

1) VALUE OF INDIVIDUAL FEATURES

We computed Pearson correlation coefficients [38] between each of the speech features and each of the child's characteristics (age, total ADOS score, SA score, and RRB score). This revealed potential relationships between the magnitude of each feature (e.g., pitch variability) and the severity of autism symptoms.

2) PREDICTION OF ADOS SCORES

To predict ADOS scores we trained and tested the models described above on independent samples. We divided the 72 available recordings into a training dataset with 51 children (70%) and a testing dataset with 21 children (30%). To ensure that both the train and test datasets retained the distribution of ADOS scores in the initial data we implemented a balanced cross validation procedure, by randomly creating train and test groups that fulfilled the following criteria:

- Mean ADOS score of each group had to be within the range of -10% and $+10\%$ of the total sample mean.
- The standard deviation of ADOS scores in each group had to be within the range of -10% and $+10\%$ of the total sample standard deviation.
- Kurtosis of the ADOS score distribution of each group had to be within the range of -10% and $+10\%$ of the kurtosis of the total sample distribution.
- Skewness of the distribution of each group had to be between -0.3 and $+0.3$.

If one of the conditions was not met, we randomly selected another pair. We created 50 different train and test groups and tested each of the 6 models on all of them to demonstrate the generalizability of the findings across different data selections.

The prediction accuracy of each model was assessed for each of the 50 datasets by computing the RMSE and Pearson's correlation coefficient between the true ADOS scores and the ADOS scores predicted by the model.

IV. RESULTS

We identified a total of 27,395 vocal segments in the recordings of the 72 children, which were used in the following analyses:

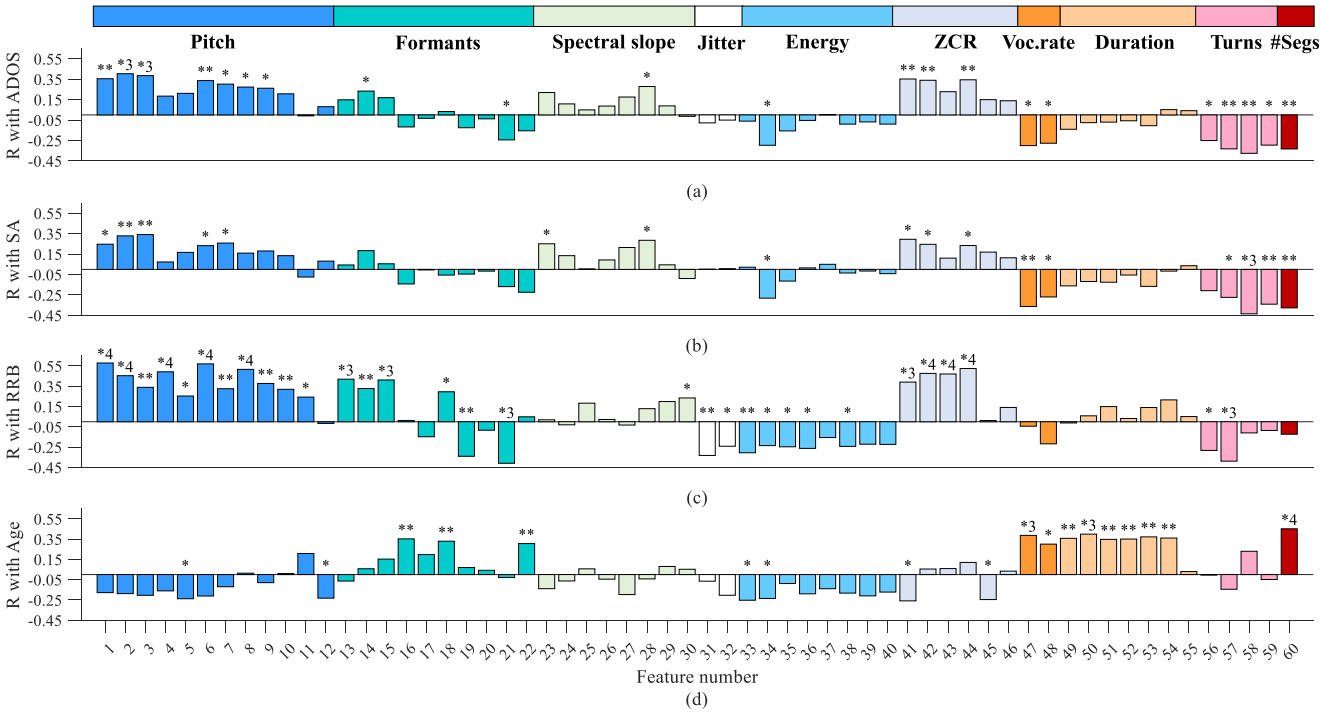

FIGURE 5. Correlation coefficients between the feature set and children characteristics. (a) Features correlations with ADOS score. (b) Features correlations with ADOS-SA score. (c) Features correlations with ADOS-RRB score. (d) Features correlations with age. Each color represents different feature group. Asterisks indicate significant correlations ($*p$-value < 0.05, $**$ p-value ≤ 0.01, $*3$ p-value ≤ 0.001, $*4$ p-value ≤ 0.0001).

A. VALUE OF INDIVIDUAL FEATURES

Twenty-one out of the 60 examined features were significantly correlated with the ADOS scores of the examined children (Fig. 5). Seventeen features were significantly correlated with the ADOS SA scores, 31 features were correlated with the RRB scores, and 18 features were correlated with the age of the children. Note that there was little overlap between the speech features that were correlated with age and those that were correlated with ASD severity, indicating that distinct features carry information about these different child characteristics.

B. ADOS ESTIMATION

We initially compared the performance of all six models using a single training (i.e., 51 recordings) and testing (i.e., 21 recordings) datasets. The behavioral characteristics and the age of the children in the training and testing datasets were intentionally matched (Table 4).

1) MULTIPLE LINEAR REGRESSION

We performed the multiple linear regression analysis with the 15 most informative features (#6, 9, 10, 13, 18-21, 24, 26, 27, 29, 32, 40 and 50), as selected by the FS procedure described above (Fig. 3). The predicted ADOS scores (rounded to the closest number) were moderately correlated with the true ADOS scores (R = 0.43, p-value = 0.05) with an RMSE of 6.93 (Fig. 6a).

TABLE 4. Children characteristics in initial train and test dataset.

Dataset	Size	ADOS (μ±σ)	Age (μ±σ)	SA (μ±σ)	RRB (μ±σ)
Train	51	12.1±6.6	51.9±14.2	8.8±5.4	3.5±2.0
Test	21	12.1±6.9	45.6±17.4	8.7±5.5	3.4±2.0

2) SVR

An SVR model trained with all 60 features, yielded predicted ADOS scores that were strongly and significantly correlated with the actual ADOS scores (R = 0.78, p-value < 0.0001), but with a relatively high RMSE of 5.56 (Fig. 6b). An SVR model trained with the 15 selected features in the multiple linear regression yielded poorer results (R = 0.42, p-value > 0.05, RMSE = 6.45).

3) DNN MODELS

There were considerable differences in the performance of the four examined DNN models (Fig. 6). The CNN model yielded the highest correlation (R = 0.82, p-value < 0.0001) and lowest RMSE (3.83) of all models. FC-DNN model 3 yielded a slightly weaker correlation (R = 0.81, p-value < 0.0001) and higher RMSE (3.97). FC-DNN model 1 yielded a weaker correlation (R = 0.77, p-value < 0.0001) and higher RMSE (4.49), and FC-DNN model 2 yielded the lowest correlation (R = 0.47, p-value < 0.05) and highest RMSE (6.12).

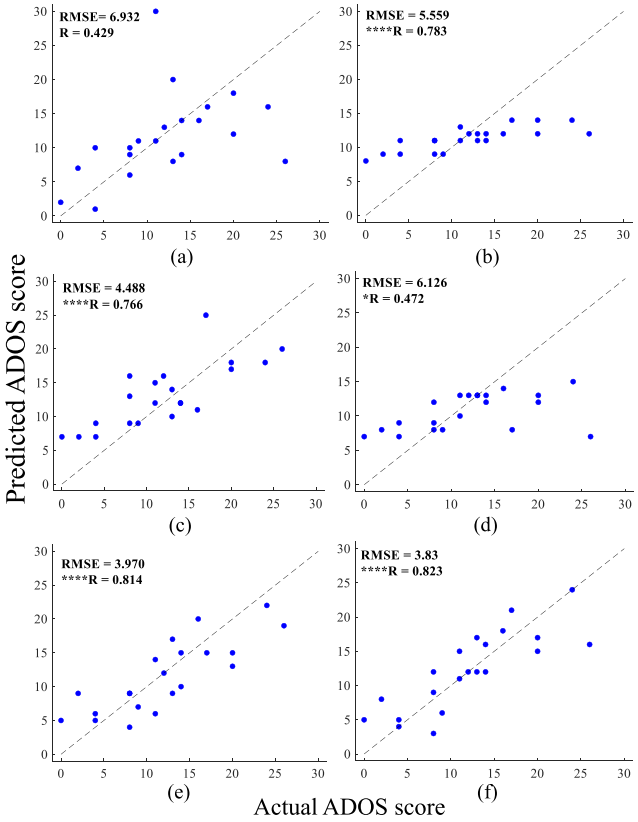

FIGURE 6. Scatter plots of the predicted ADOS scores vs. actual ADOS scores for each of the 6 models when using one selection of training and testing datasets. (a) Linear regression. (b) SVR model. (c) FC-DNN model 1 (d) FC-DNN model 2. (e) FC-DNN model 3. (f) CNN model. The RMSE and correlation coefficient (R) are presented in each panel. $*p$-value < 0.05, $****p$-value < 0.0001.

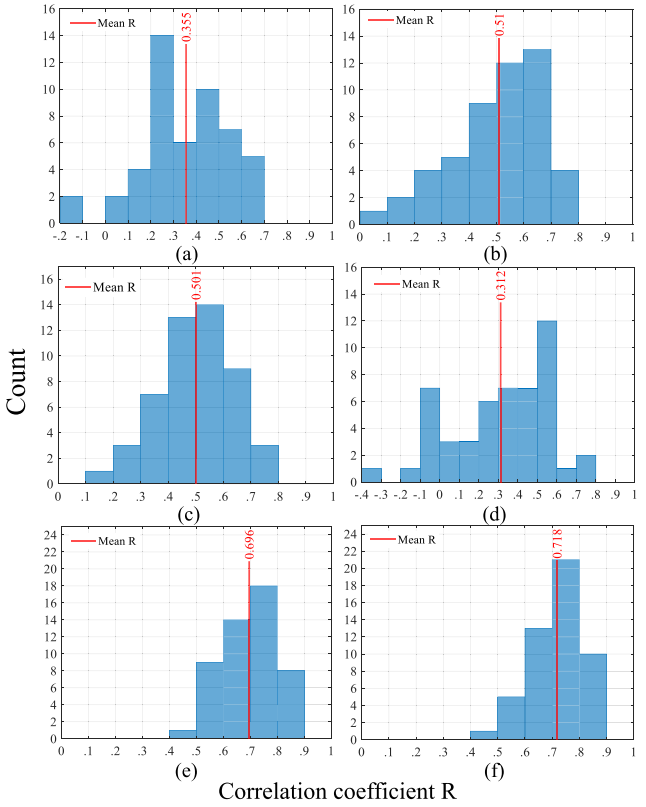

FIGURE 7. Histograms demonstrating the distribution of R values for each of the 6 models when tested on 50 different selections of balanced testing and training datasets. (a) Linear regression model. R = 0.355 ± 0.193, RMSE = 7.364 ± 1.298. (b) SVR model. R = 0.510 ± 0.166, RMSE = 5.853 ± 0.431. (c) FC-DNN model 1. R = 0.501 ± 0.134, RMSE = 6.148 ± 0.864. (d) FC-DNN model 2. R = 0.312 ± 0.253, RMSE = 6.393 ± 0.701. (e) FC-DNN model 3. R = 0.696 ± 0.091, RMSE = 4.952 ± 0.651. (f) CNN model. R = 0.718 ± 0.093, RMSE = 4.648 ± 0.592.

4) BALANCED CROSS VALIDATION

To further evaluate the generalizability of these findings to alternative selections of training and testing datasets, we randomly selected 50 balanced training and testing datasets (as described above) and re-tested each of the six models (two regression and 4 DNN models) with each selection. This yielded a histogram summarizing the performance of each model across the 50 selections (Fig. 7).

The CNN model demonstrated the best performance across datasets with the highest correlations (mean R = 0.72 ± 0.09) and lowest RMSE values (mean = 4.65 ± 0.59). FC-DNN model 3 followed with lower correlations (mean R = 0.70 ± 0.09) and higher RMSE values (mean = 4.95 ± 0.65). Then the SVR model (mean R = 0.51 ± 0.17 and mean RMSE = 5.86 ± 0.43), FC-DNN model 1 (mean R = 0.50 ± 0.13 and mean RMSE = 6.15 ± 0.86), and the linear regression model (mean R = 0.36 ± 0.19 and mean RMSE = 7.36 ± 1.3). Finally, FC-DNN model 2 exhibited the poorest performance (mean R = 0.31 ± 0.25 and mean RMSE = 6.39 ± 0.7).

Note that the variability of performance across datasets differed across models. The CNN model and FC-DNN model 3 exhibited the most consistent performance across the 50 dataset selections such that the standard deviation of correlation values for both models was 0.09. In contrast, the other models were considerably less consistent in their performance across dataset selections, exhibiting large standard deviations that were between 0.13 – 0.25. These results demonstrate the importance of selecting different training and testing samples for determining the robustness and consistency of performance.

V. DISCUSSION

Our study demonstrates that DNN models with specific architectures can be trained to predict ASD severity from speech recordings of children with remarkably high accuracy. When using the CNN model or FC-DNN model 3, the predicted ADOS scores were strongly and consistently correlated with the actual ADOS scores reported by the clinician (Fig. 7). The models were trained with values of specific speech features that were extracted from recordings of clinical ADOS assessments where each child with ASD interacted with a clinician for ∼40 min. While previous studies have attempted to use such recordings for separating ASD and typically developing children [39], only a few studies to date have utilized these speech features to predict the actual severity of

ASD symptoms (i.e., ADOS scores). Furthermore, this is the first time that these speech features and deep learning technique have been applied to recordings of Hebrew speaking children. These results highlight the utility of speech analyses for estimating ASD severity at very young ages, regardless of the child's spoken language or their cultural environment. Further development of these algorithms has great potential for aiding clinicians in assessing early risk for ASD and for quantifying changes in ASD severity over time and in response to treatment.

A. VALUE OF INDIVIDUAL FEATURES

The results revealed that specific prosodic, acoustic, and conversational features from individual recordings were significantly correlated with their ADOS scores (Fig. 5). Positive correlations were apparent with most of the pitch (1-12) and ZCR (41-46) features indicating that Hebrew speaking ASD children with more severe symptoms tend to speak with higher pitch, larger pitch variability, higher rates of zero crossing values, and larger variability of zero crossing values. These results are in line with previous studies demonstrating that ASD children speak with larger pitch variability in comparison to controls. This was true for both Hebrew [40] and English [41], [42] speaking children.

In contrast, conversational features of turn-taking (56-59) and the total number of spoken segments (60) were negatively correlated with ADOS scores. This indicates that ASD children with more severe symptoms speak less, participate in fewer conversational turns, as also reported in previous studies with English speaking children [43]. Surprisingly, our results showed that when ASD children with more severe symptoms did respond, they tended to do so more quickly (feature #56). This contradicts previous studies showing the opposite in older English speaking children [43]. Further research is necessary to assess the reproducibility of either finding.

The magnitude of correlations with specific features of speech differed when separating the total ADOS scores into their SA and RRB components. RRB scores were more strongly correlated with pitch, formants, jitter, energy, and zero crossing features, while SA scores were more correlated with vocalization rate, turn-taking, and total number of vocalizations. This suggests that RRB symptoms tend to be more strongly associated with acoustic and prosodic features while SA symptoms tend to be more strongly associated with conversational features. The specificity of particular speech features to specific ASD symptom domains is critical for developing speech analysis tools with clinical utility, because children with ASD exhibit heterogeneous symptoms that can differ dramatically across ASD cases.

Another important finding is the clear dissociation between features that were associated with ASD severity (ADOS scores) and features that were associated with the age of the children. The age of participating children was most strongly correlated with only few extracted features: vocalization rate, the duration of vocalizations, and the total number of vocalizations. Most acoustic and prosodic features were not significantly correlated with age. Note that the utility of specific speech features for assessment of ASD severity may change with age, requiring a longitudinal research approach that tests the utility of different speech features during multiple developmental periods.

B. ADOS SEVERITY ESTIMATION

The speech features described above were used to train the six models that were built to estimate ADOS severity. There were considerable differences in the performance of the six models such that the CNN model and FC-DNN model 3 outperformed the four other models in a robust and consistent manner (Fig. 7).

The six models differed in their architecture and the structure of their input data. In FC-DNN model 1 we used each recording as a single sample while computing a single value for each speech feature across the entire recording (i.e., single input vector with the values of 60 features per recording/child). The performance of this model was better than that of a linear regression model and slightly worse than the SVR model.

In an attempt to improve ADOS prediction we created FC-DNN model 2 where we changed the input data such that instead of having a single sample per recording/child, we now extracted 100 random samples of 10 consecutive vocal segments from each recording. We computed 48 of the 60 speech features for each of these samples/vectors, thereby yielding 100 samples per recording/child instead of just one. This new architecture utilized the fact that there were many child vocalizations in each recording, enabling us to create multiple samples from each recording. The disadvantage of this approach was that we had to limit the feature vector to the 48 acoustic and prosodic features that could be computed from child vocalizations only, without the 12 conversational features that require assessment of the entire recording. This approach, however, did not work well, yielding the poorest performance of all models.

In FC-DNN model 3 we combined the same 100 randomly selected samples of 10 vocal segments into a single input matrix (size: 100×48). This approach yielded considerably better performance in comparison to all other FC-DNN and regression models, despite the limitation of utilizing only 48 acoustic and prosodic features (i.e., without the conversational features).

Finally, altering the architecture to a CNN model while using the same input structure as FC-DNN model 3, yielded an additional increase in performance. This suggests that there is a large performance benefit to architectures that take advantage of input containing multiple samples of the recorded child's speech (i.e., matrix with speech features as extracted from multiple combinations of speech segments). Furthermore, the considerable improvement in performance of the CNN model and FC-DNN model 3, relative to the linear regression and support vector regression models,

demonstrates that the optimal ADOS prediction model is not likely to be linear.

Previous studies in this domain have mostly focused on utilizing speech features to classify children into ASD and typically developing groups. For example, Pokorny *et al.* [39] examined recordings of vocalizations from twenty 10-month-old children later diagnosed with ASD, and a matched TD group. They extracted 88 acoustic features of the children's vocalizations and achieved 75% accuracy in classifying children into their respective groups using a BLSTM DNN algorithm.

To the best of our knowledge, only a few studies to date have attempted to estimate ASD severity from vocal recordings of children. For example, one study [14] used a combined DNN and Random Forest algorithm to estimate Calibrated Severity SA Scores (CSS SA) from 33 recordings and was able to predict CSS SA scores yielding a correlation of ∼0.63 with the true CSS SA scores. Their approach utilized a DNN for speech activity detection and speaker diarization, and a synthetic random forest algorithm for ADOS score estimation.

Further research incorporating automatic speaker diarization techniques, instead of manual annotation, will enable development of fully automated ASD severity estimation systems. Furthermore, additional research will be able to extend these techniques for use with longer home recordings as performed using the LENA system [11] rather than short recordings performed at the clinic.

VI. CONCLUSION

Our results demonstrate that a variety of prosodic, acoustic, and conversational features are informative of ASD severity in young Hebrew speaking children. These features can be utilized by a CNN model to yield remarkably accurate prediction of ADOS scores when applying an architecture that utilizes multiple vocalization samples from each child in tandem. We suggest that this speech analysis algorithm may have considerable clinical utility in assessing early ASD risk and as a novel outcome measure for quantifying ASD severity changes over time and following treatments.

ACKNOWLEDGMENT

The authors would like to thank Alex Gorodetski, Carmel Most, Reut Altman, and Liat Moyal for their support in performing this research. They would also like to thank the Israeli Ministry of Science and Technology for creating and funding the National Autism Research Center of Israel, which enabled the collection of the data analyzed in this study.

REFERENCES

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *Lancet*, vol. 392, no. 10146, pp. 508–520, 2018.
[2] H. Tager-Flusberg, "Defining language phenotypes in autism," *Clin. Neurosci. Res.*, vol. 6, nos. 3–4, pp. 219–224, Oct. 2006.
[3] I. Rapin and M. Dunn, "Update on the language disorders of individuals on the autistic spectrum," *Brain Develop.*, vol. 25, no. 3, pp. 166–172, Apr. 2003.
[4] M. Mergl, C. Alves, and S. Azoni, "Echolalia's types in children with Autism spectrum disorder," *Revista CEFAC*, vol. 17, no. 6, pp. 2072–2080, 2015.
[5] A. Shield, F. Cooley, and R. P. Meier, "Sign language echolalia in deaf children with autism spectrum disorder," *J. Speech, Lang., Hearing Res.*, vol. 60, no. 6, pp. 1622–1634, Jun. 2017.
[6] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: A critical review," *Int. J. Lang. Commun. Disorders*, vol. 38, no. 4, pp. 325–350, Jan. 2003.
[7] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "'Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis,'" *Autism Res.*, vol. 10, no. 3, pp. 384–407, 2017.
[8] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech, Lang., Hearing Res.*, vol. 57, no. 4, pp. 1162–1177, Aug. 2014.
[9] R. B. Grossman, R. H. Bemis, D. Plesa Skwerer, and H. Tager-Flusberg, "Lexical and affective prosody in children with high-functioning autism," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 3, pp. 778–793, Jun. 2010.
[10] L. Schreibman, G. Dawson, A. C. Stahmer, R. Landa, S. J. Rogers, G. G. McGee, C. Kasari, B. Ingersoll, A. P. Kaiser, Y. Bruinsma, E. McNerney, A. Wetherby, and A. Halladay, "Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder," *J. Autism Develop. Disorders*, vol. 45, no. 8, pp. 2411–2428, Aug. 2015.
[11] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 30, pp. 13354–13359, Jul. 2010.
[12] J. Rankine, E. Li, S. Lurie, H. Rieger, E. Fourie, P. M. Siper, A. T. Wang, J. D. Buxbaum, and A. Kolevzon, "Language ENvironment analysis (LENA) in phelan-McDermid syndrome: Validity and suggestions for use in minimally verbal children with autism spectrum disorder," *J. Autism Develop. Disorders*, vol. 47, no. 6, pp. 1605–1617, Jun. 2017.
[13] A. L. Harbison, T. G. Woynaroski, J. Tapp, J. W. Wade, A. S. Warlaumont, and P. J. Yoder, "A new measure of child vocal reciprocity in children with autism spectrum disorder," *Autism Res.*, vol. 11, no. 6, pp. 903–915, Jun. 2018.
[14] S. Sadiq, M. Castellanos, J. Moffitt, M.-L. Shyu, L. Perry, and D. Messinger, "Deep learning based multimedia data mining for autism spectrum disorder (ASD) diagnosis," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 847–854.
[15] *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. American Psychiatric Association, Arlington, TX, USA, 2013.
[16] K. Gotham, A. Pickles, and C. Lord, "Standardizing ADOS scores for a measure of severity in autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 39, no. 5, pp. 693–705, May 2009.
[17] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT," *Glot Int.*, vol. 5, nos. 9–10, pp. 341–347, 2001.
[18] S. Schelinski and K. von Kriegstein, "The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development," *J. Autism Develop. Disorders*, vol. 49, no. 1, pp. 68–82, Jan. 2019.
[19] P. Jongmans, T. G. Wempe, H. van Tinteren, F. J. M. Hilgers, L. C. W. Pols, and C. J. van As-Brooks, "Acoustic analysis of the voiced-voiceless distinction in dutch tracheoesophageal speech," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 2, pp. 284–297, Apr. 2010.
[20] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Appl. Soft Comput.*, vol. 27, pp. 231–239, Feb. 2015.
[21] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *J. Autism Develop. Disorders*, vol. 44, no. 10, pp. 2413–2428, Oct. 2014.
[22] L. Tamarit, M. Goudbeek, and K. Scherer, "Spectral slope measurements in emotionally expressive speech," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Anal. Process. for Knowl. Discovery*, 2008, pp. 1–4, Paper 007.

[23] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[24] M. Guzman, S. Correa, D. Muñoz, and R. Mayerhoff, "Influence on spectral energy distribution of emotional expression," *J. Voice*, vol. 27, no. 1, pp. 129.e1-129.e10, 2013.
[25] F. Alías, J. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Appl. Sci.*, vol. 6, no. 5, p. 143, May 2016.
[26] L. G. Pillai and E. Sherly, "A deep learning based evaluation of articulation disorder and learning assistive system for autistic children," *Int. J. Natural Lang. Comput.*, vol. 6, no. 5, pp. 19–36, Oct. 2017.
[27] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 10, pp. 5132–5148, 2017.
[28] M. Kriboy, A. Tarasiuk, and Y. Zigel, "A novel method for obstructive sleep apnea severity estimation using speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3606–3610.
[29] S. Gillespie, E. Moore, J. Laures-Gore, M. Farina, S. Russell, and Y.-Y. Logan, "Detecting stress and depression in adults with aphasia through speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5140–5144.
[30] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5005–5009.
[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[32] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Comput. Speech Lang.*, vol. 56, pp. 80–94, Jul. 2019.
[33] S. Gupta, K. De, D. A. Dinesh, and V. Thenkanidiyoor, "Emotion recognition from varying length patterns of speech using CNN-based segment-level pyramid match kernel based SVMs," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2019, pp. 1–6.
[34] S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.
[35] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Netw.*, vol. 64, pp. 39–48, Apr. 2015.
[36] J. Lyu and S. Sheen, "A channel-pruned and weight-binarized convolutional neural network for keyword spotting," in *Proc. Int. Conf. Comput. Sci., Appl. Math. Appl.*, 2019, pp. 243–254.
[37] 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," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 140–145.
[38] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 757–765, May 2008.
[39] F. B. Pokorny, B. Schuller, P. B. Marschik, R. Brueckner, P. Nyström, N. Cummins, S. Bölte, C. Einspieler, and T. Falck-Ytter, "Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach," in *Proc. Interspeech*, Aug. 2017, pp. 309–313.
[40] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal speech spectrum and increased pitch variability in young autistic children," *Frontiers Hum. Neurosci.*, vol. 4, p. 237, Jan. 2011.
[41] M. Sharda, T. P. Subhadra, S. Sahay, and C. Nagaraja, "Sounds of melody—Pitch patterns of speech in autism," *Neurosci. Lett.*, vol. 478, no. 1, pp. 42–45, 2010.
[42] L. D. Shriberg, R. Paul, L. M. Black, and J. P. van Santen, "The hypothesis of apraxia of speech in children with autism spectrum disorder," *J. Autism Develop. Disorders*, vol. 41, no. 4, pp. 405–426, Apr. 2011.
[43] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. Narayanan, "Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders," *Interspeech*, vol. 2016, pp. 1185–1189, Sep. 2016.

MARINA ENI received the B.Sc. degree in biomedical engineering from the Ben-Gurion University of the Negev, Beersheba, Israel, in 2018, where she is currently pursuing the master's degree. Her current research interests include audio and image processing, biomedical engineering, pattern recognition, and artificial intelligence.

ILAN DINSTEIN received the B.Sc. degree in life sciences from Tel Aviv University and the Ph.D. degree in neuroscience from New York University. He completed Postdoctoral Training with the Weizmann Institute of Science and Carnegie Mellon University. In 2012, he joined the Psychology Department, Ben-Gurion University of the Negev, where he is currently an Associate Professor and the Director of the National Autism Research Center of Israel.

MICHAL ILAN received the M.A. degree in linguistics from Bar-Ilan University, in 2008. She is currently pursuing the Ph.D. degree with the Psychology Department, Ben-Gurion University of the Negev, studying the effect of educational settings on changes in Autism symptoms, in preschool children. She has been a Speech and Language Therapist, since 2005. She is specializes in working with children with Autism. She is currently working with the Soroka University Medical Center, ASD Team, Pre-School Psychiatric Unit.

IDAN MENASHE received the B.Sc. degree in medical sciences from Ben-Gurion University and the M.Sc. and Ph.D. degrees in human genetics and bioinformatics from the Weizmann Institute of Science. He also had a Postdoctoral Training in biostatistics with the Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute (NCI), USA, and then worked as a Senior Scientist with MindSpec Inc. In 2012, he joined the Public Health Department, Ben-Gurion University (BGU). His main research interest includes genetic epidemiology, where he is studying the effect of genetic and non-genetic factors on the risk of autism and other complex disorders. He also serves as the Scientific Director of the National Autism Research Center of Israel.

GAL MEIRI received the M.D. and M.H.A. degrees from the Ben-Gurion University of the Negev. He did part of his residency and a Research Fellowship with the Oregon Health Sciences University at Portland, Oregon. He is currently a Child Psychiatrist and the Head of the Soroka University Medical Center, Preschool Psychiatric Unit. He is an Associate Professor with the Faculty of Health Sciences, Ben-Gurion University of the Negev. He is involved in the community and serves in local and national committees. He was recently the President of the Israeli Association of Infant Mental Health (Affiliated with the WAIMH). He serves as an Active Member of the National Committee of Experts on Autism in the Israeli Ministry of Health. He and his partners from Soroka and BGU established in 2014 the Negev Autism Center and in 2018 this center was announced by the Ministry of Sciences as the National Autism Research Center of Israel.

YANIV ZIGEL (Senior Member, IEEE) was born in Tel-Aviv, Israel, in 1970. He received the B.Sc., M.Sc., *(summa cum laude)*, and Ph.D. degrees in electrical and computer engineering from Ben-Gurion University, Beersheba, Israel, in 1992, 1998, and 2004, respectively. His M.Sc. thesis and Ph.D. research were in the fields of biomedical and speech signal processing. From 2003 to 2006, he held a position as a Senior DSP Algorithm Engineer with the Audio Analysis Group, NICE Systems Ltd. From 2006 to 2007, he was the Speech Research Group Leader, PuddingMedia Ltd. Since 2007, he has been a Faculty Member with the Department of Biomedical Engineering, Faculty of Engineering Sciences, Ben-Gurion University of the Negev, and the Head of the Biomedical Signal Processing Research Laboratory. His main interests include biomedical signal processing, speech and audio analysis, and pattern recognition.