The Impact of Speaker Diarization on DNN-based Autism Severity Estimation*

Marina Eni, Alex Gorodetski, Ilan Dinstein, and Yaniv Zigel, *Senior member, IEEE*

***Abstract*— This paper presents a speech-based system for autism severity estimation combined with automatic speaker diarization. Speaker diarization was performed by two different methods. The first used acoustic features, which included Mel-Frequency Cepstral Coefficients (MFCC) and pitch, and the second used x-vectors – embeddings extracted from Deep Neural Networks (DNN). The speaker diarization was trained using a Fully Connected Deep Neural Network (FCDNN) in both methods. We then trained a Convolutional Neural Network (CNN) to estimate the severity of autism based on 48 acoustic and prosodic features of speech. One hundred thirty-two young children were recorded in the Autism Diagnostic Observation Schedule (ADOS) examination room, using a distant microphone. Between the two diarization methods, the MFCC and Pitch achieved a better Diarization Error Rate (DER) of 26.91%. Using this diarization method, the severity estimation system achieved a correlation of 0.606 (Pearson) between the predicted and the actual autism severity scores (i.e., ADOS scores).**

***Clinical Relevance*—The presented system identifies children's speech segments and estimates their autism severity score.**

I. Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder diagnosed by social communication impairments, repetitive behaviors, and confined interests [1]. A common clinical tool for assessing ASD severity is the Autism Diagnostic Observation Schedule (ADOS) [2], a 45-min assessment where a clinician interacts with the child using a variety of tasks/games and scores the child's behavior using a standardized scale. Raw ADOS scores range between 0 and 30, with higher values indicating more severe ASD symptoms.

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 [3]. Previous studies have demonstrated that verbal children with ASD speak in abnormal ways, including strange prosody [4], increased pitch variability and pitch range [5], slower speech rate, and prolonged word production. These studies used manual annotation and inspection of the audio recordings.

To perform automatic analysis of ADOS recordings, there is a need for a speaker diarization algorithm that divides the recording into segments of child and of other speakers. Speaker diarization is affected by the quality of the audio recording, the number of recording microphones, the distance between microphones and speaking subjects. This task is especially challenging in real-life recordings, like sessions performed in a clinic during the ADOS diagnosis, including background noise, reverberation, and overlapping speech [6].

Recent studies have utilized automated speech processing algorithms to examine speech recordings of children with ASD. Deep Neural Network (DNN) has been previously used for speaker diarization, such as Long Short-Term Memory (LSTM) [7] and Region Proposal Network (RPN) [8] networks. Most recent studies combined DNNs with clustering approaches, such as k-means [9] or Probabilistic Linear Discriminant Analysis (PLDA) [10] to improve their speaker diarization performance.

In recent years, speech features have been used in autism detection, classifying children/adults into ASD or typically developed (TD) groups [11]. While these studies reported high classification performance, they did not estimate the severity of autism symptoms in individual children. Nevertheless, one study has conducted an estimation of Japanese-speaking adolescents' social abilities using a classical machine learning classifier, Support Vector Regression (SVR) [12]. Additional work was done on English-speaking children, which also predicted the social skills of the children with autism using a DNN framework [13].

This paper presents a speech-based system for autism severity estimation of Hebrew-speaking children, integrated with an automatic speaker diarization. Two diarization methods were compared and their impact on the estimation system was examined.

II. Experimental Setup

Audio recordings of 132 children were acquired from ADOS-2 assessments that were performed at the Azrieli National Centre for Autism and Neurodevelopment Research (ANCAN) by a trained clinician with ADOS-2 research reliability. The recordings were performed with a single microphone (CHM99, AKG, Vienna), located 1-2 m from the child (the child and the therapist can move during the session). Each ADOS session was recorded at a sampling rate of 44.1 kHz (downsampled to 16 kHz), and the mean duration of each session was 41.96 ± 11.51min. Of the 132 children included in the study, 105 were diagnosed with ASD, and the others (non-

*Research supported by Israel Ministry of Science and Technology (Grant no. 3-17422).

M. Eni, A. Gorodetski, I. Dinstein, and Y. Zigel are with the Ben-Gurion University of the Negev, Beer-Sheva, 8410501 Israel (e-mail: marinamu@post.bgu.ac.il).

ASD) were diagnosed with language or developmental delays or were defined as typically developing, see Table I.

TABLE I. CHARACTERISTICS OF PARTICIPATING CHILDREN INCLUDING THEIR ASD SEVERITY AS ESTIMATED BY THE ADOS. MEAN (STD.)

Group	N	Age (y)	ADOS score	Male (%)
ASD	105	3.7 (1.3)	14.6 (6.5)	80.0
Non-ASD	27	3.1 (1.5)	3.2 (3.7)	77.8

III. METHODS

The presented system, see Fig. 1, includes a speaker diarization algorithm, which divides the audio recording into child speech segments and other speakers or audio events (e.g., movements). Next, prosodic and acoustic features are extracted from the detected child segments and are then used as an input for the ASD severity estimation model.

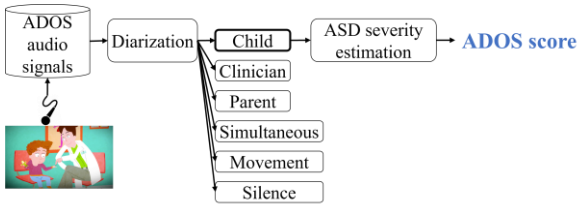

Figure 1. A simplified block diagram of the ASD severity estimation system combined with speaker diarization.

A. Manual labeling

The audio recordings were manually labeled using an in-house software. Audio segments were labeled as child, therapist, parent, movement, or simultaneous speech (i.e., a speech of more than one speaker). All remaining segments were automatically labeled as silence.

B. Diarization

Speaker diarization was performed by two different methods:

1. Set of 13 Mel-Frequency Cepstral Coefficients (MFCCs), pitch (fundamental frequency), and their first derivatives, total of 28 features.
2. An x-vector consisting of a set of 512 features.

We applied the same DNN architecture to each of these feature sets. The DNN included a Fully Connected (FC) network (FCDNN, see Fig. 2), involving FC layers, which were followed by two Softmax layers (number of outputs = number of classes); the second Softmax layer was added to improve the final classification [14]. We performed these analyses while classifying speech segments into six classes: child, therapist, parent, movement, simultaneous, and silence.

The x-vector features were calculated using a system developed by the Brno University of Technology (BUT) [15] with a window size of 1.5s and a window rate of 10ms. Then, a Viterbi algorithm [16] was applied to the output of the FCDNN. Finally, we used only the child-detected segments for further analysis.

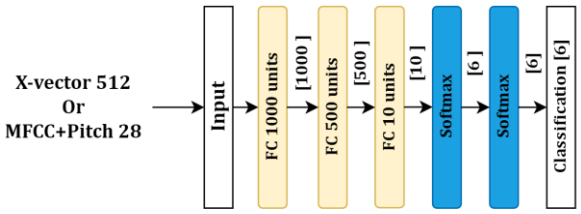

Figure 3. The architecture of the FCDNN for the speaker diarization.

C. Autism severity estimation

The detected/manually labeled child segments often contained multiple vocalizations (e.g., multiple utterances) separated by silence. To separate them, we used our previously reported method [17], where now we also included the long vocalizations (longer than 3s). Next, for each recording, we extracted a feature matrix of size 100×48 [17]. Each row corresponds to a feature vector of 48 prosodic and acoustic features (such as pitch, formants, Zero Crossing Rate, duration, etc.), which was calculated for a random set of 10 sequential vocalizations of the child.

The estimation of the autism severity (i.e., ADOS scores) was based on a CNN architecture (Fig. 3). For a detailed explanation see [17]. Here, we used the same architecture, for training and testing the system with manually labeled child vocalizations (baseline method), and again for training and testing the system with the output of the proposed diarization method. In both cases, we trained the networks using the RMSprop (Root Mean Square Propagation) optimizer [18] with a linear output (regression), and parameters that were tuned using a random search algorithm with a 5-fold cross-validation method (see Table II).

TABLE II. ASD ESTIMATION SYSTEM HYPER-PARAMETERS.

Segmentation method		Batch size	#Epochs	LR
Baseline		16	360	1e-5
Proposed diarization methods	MFCC + Pitch	32	610	1e-5
	X-vectors	8	310	1e-5

LR stands for learning rate.

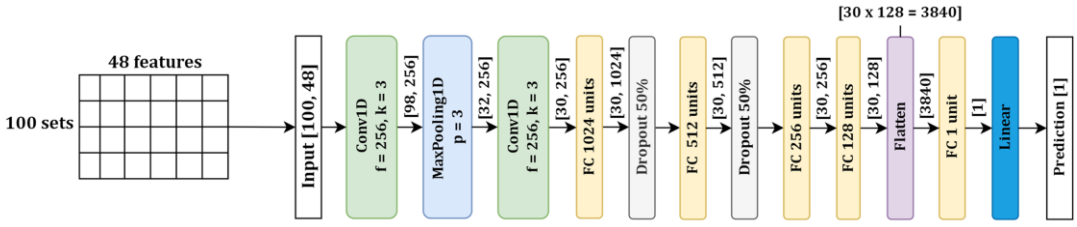

Figure 2. CNN architecture for the ASD severity estimation.

D. Performance analysis

Diarization: We evaluated the diarization model using the leave-one-out method and reported the recall, precision, accuracy, and Diarization Error Rate (DER, (1)) values. We only focused on classifying child segments and merged all other classes into a single category. DER was calculated using the dscore tool [19] with a collar value of 0.25sec [7].

$$DER = SE + FA + MD \tag{1}$$

where SE is the substitution error – the percentage of erroneously classified frames (not silence). FA is a false accept rate – the percentage of frames labeled as silence but classified as child or other. MD is the missed detection rate – the percentage of frames labeled as child or other but classified as silent. The collar defines a time segment before and after each manually labeled segment where errors are ignored, but only if the segment was classified correctly.

Since silence frames may exist between words in a child segment, and the ASD severity estimation includes removal of these frames, we calculated accuracy, recall, and precision while ignoring silence classification errors.

ASD severity estimation: To test the system's performance, it is necessary to reduce the effect of a specific data division into train and test datasets. To achieve this, we divided the children into training and testing datasets while balancing the distribution of ADOS scores (mean, standard deviation, kurtosis, and skewness) in each dataset. In this method, we maintained a similar mean, standard deviation, kurtosis, and skewness, relatively to the whole dataset. This procedure was executed 50 times with the training and testing groups consisting of 80% and 20% of the data, respectively. The system's performance was evaluated using the Normalized Root Mean Square Error (NRMSE, (2)) and the Pearson correlation coefficient R between the actual and the estimated ADOS scores.

$$NRMSE = \frac{RMSE}{\max(y_{actual}) - \min(y_{actual})} \tag{2}$$

where y_{actual} are the actual ADOS scores (taken from the train dataset) defined by the clinician.

IV. Results

Two different diarization methods were evaluated and analyzed. The total duration of the vocalizations as detected by the two diarization methods is displayed in Table III.

TABLE III. THE TOTAL DURATION OF DIFFERENT DIARIZATION METHODS

	Baseline	**MFCC + Pitch**	**X-vectors**
Total duration [h]	10.85	11.11	10.72

Fig. 4 provides the results of both speaker diarization methods in separating child segments from all other categories. The figure shows that the detected child segments were effectively classified using the x-vectors, with the recall, precision, and accuracy of 73.42%, 74.06%, and 89.84%, respectively. The MFCC+Pitch method derived the lowest classification error of 26.91% in terms of DER. Figure 5 shows an example of the output of the diarization methods on a short audio signal.

The correlation between actual and predicted ADOS scores was higher when using the diarization methods than manual labeling (see Table IV). The Pearson correlation coefficient was higher and NRMSE was lower when using child vocalization identified by the MFCC+Pitch diarization method than when using manually labeled vocalizations.

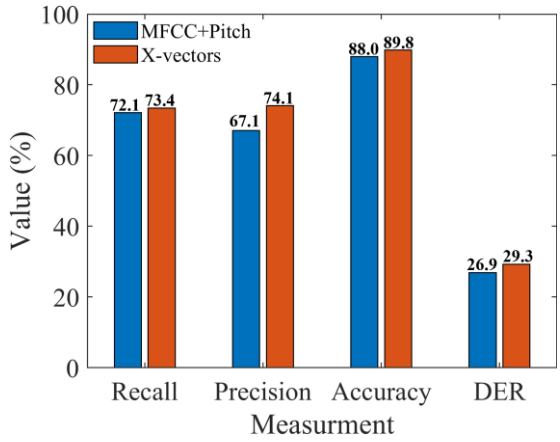

Figure 4. Performance of the two diarization methods in identifying child speech segments.

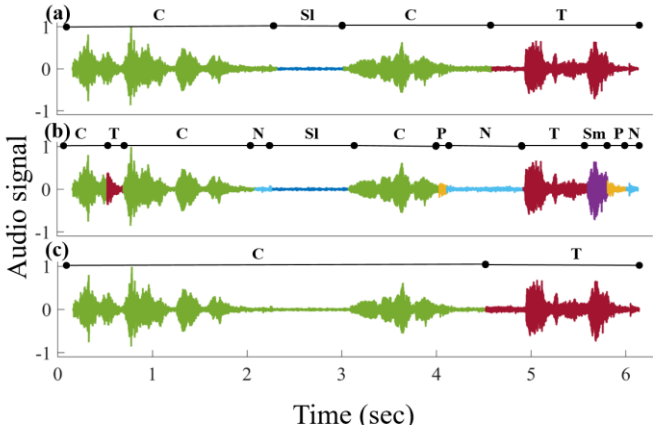

Figure 5. Example of the three segmentation methods on a short audio signal: (**a**) the baseline segmentation (manually labeled), (**b**) MFCC+Pitch diarization method, and (**c**) X-vectors diarization method. The different colors represent different detected speakers: child (green, C), therapist (dark red, T), noise (light blue, N), silence (dark blue, Sl), parent (orange, P), and simultaneous (purple, Sm).

TABLE IV. ASD SEVERITY ESTIMATION PERFORMANCE RESULTS OF PREDICTED VERSUS ACTUAL ADOS SCORES WHEN USING MANUAL ANNOTATION OR EACH OF THE TWO DIARIZATION METHODS: MEAN ± STD.

Segmentation method	**NRMSE**	**Pearson R**
Baseline	0.261 ± 0.036	0.567 ± 0.120
MFCC + Pitch	**0.248 ± 0.028**	**0.606 ± 0.099**
X-vectors	0.254 ± 0.032	0.581 ± 0.115

The best result is marked in bold.

V. Discussion

The impact of speaker diarization on a DNN-based autism severity estimation was explored. Two speaker diarization methods were presented and compared, where the MFCC+Pitch showed lower DER values while the x-vector method showed slightly higher accuracy performance.

Our ASD severity estimation system, integrating automatic diarization, yielded equivalent or slightly better results than estimation based on manual segmentation. This is a very encouraging result and may be explained by the fact that the manual segmentation is not perfectly accurate, especially for the child segments. The child speech is often quiet and is sometimes masked by noises, especially when using a distant microphone in a reverberant environment. This audio quality (average SNR of child speech was 12 ± 5dB) can be considered as one limitation of this study. In addition, since it is a DNN-based system, which benefits from a large dataset, it needs a minimum duration of child speech; this was set to 20s (at least 50 vocalizations). This can be considered as a second limitation of this study.

Several diarization algorithms have been proposed in the past, where in recent years there has been a growing interest in analyzing children's speech, particularly the speech of children with autism [9], [10]. M. Pal et al. [9] applied a diarization method for speech signals using four microphones, including a beamforming algorithm that improved their recordings' SNR. Even though our proposed diarization methods achieved higher DER values, we used only one distant microphone recorded in a noisy environment. Additional study used a large ADOS corpus (499 hours of child speech) of verbally fluent children and a pitch variation augmentation method to increase their performance [10]. In comparison, our study used ~11 hours of vocalizations of children that mostly do not speak fluently or use little or no phrase speech.

Our presented system used a large sample set of young (1-7y) and not verbally fluent children. To date, only a few studies have attempted to assess the severity of autism of children from speech. Of these, one study estimated a severity score of social skills in adolescents and adults who can speak fluently and create complex conversations [12]. An additional study estimated the calibrated severity score of the social skills of children with ASD [13]. They applied a CNN model for detecting speech-related sounds and a Synthetic Random Forest for ASD severity estimation. Estimating the ASD severity of 33 children, they achieved a correlation of 0.634. In our study, we achieved a similar correlation value (0.606) while using a larger dataset of 132 children.

VI. Conclusion

This paper presents a speech-based system for autism severity estimation combined with automatic speaker diarization. At the diarization stage, the MFCC+Pitch achieved a better DER performance. The ASD severity estimation system achieved good and similar results when using either automatic diarization or manual segmentation. The current study demonstrates the utility of an end-to-end system for estimating ASD severity that integrates a speaker diarization algorithm.

References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *Lancet*, vol. 392, no. 10146, pp. 508–520, Aug. 2018.

[2] V. Hus, K. Gotham, and C. Lord, "Standardizing ADOS Domain Scores: Separating Severity of Social Affect and Restricted and Repetitive Behaviors," *J. Autism Dev. Disord.*, vol. 44, no. 10, pp. 2400–2412, Oct. 2014.

[3] H. Tager-Flusberg, "Defining language phenotypes in autism," *Clin. Neurosci. Res.*, vol. 6, no. 3–4, pp. 219–224, Oct. 2006.

[4] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: a critical review," *Int. J. Lang. Commun. Disord.*, vol. 38, no. 4, pp. 325–350, 2003.

[5] 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.

[6] M. Hu *et al.*, "Single-channel speaker diarization based on spatial features," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[7] L. Sun *et al.*, "A Novel LSTM-Based Speech Preprocessor for Speaker Diarization in Realistic Mismatch Conditions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, vol. 2018-April, pp. 5234–5238.

[8] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, p. 101317, 2022.

[9] M. Pal *et al.*, "Speaker Diarization Using Latent Space Clustering in Generative Adversarial Network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6504–6508.

[10] S. Krishnamachari, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Developing Neural Representations for Robust Child-Adult Diarization," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 590–597.

[11] A. Mohanta, P. Mukherjee, and V. K. Mirtal, "Acoustic Features Characterization of Autism Speech for Automated Detection and Classification," in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6.

[12] M. Sakishita, C. Ogawa, K. J. Tsuchiya, T. Iwabuchi, T. Kishimoto, and Y. Kano, "Autism Spectrum Disorder's Severity Prediction Model Using Utterance Features for Automatic Diagnosis Support," in *Precision Health and Medicine*, vol. 843, 2019, pp. 83–95.

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

[14] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *ICLR*, 2017.

[15] F. Landini *et al.*, "But System for the Second Dihard Speech Diarization Challenge," *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6529–6533, 2020.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

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

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

[19] N. Ryant *et al.*, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-Septe, pp. 978–982.