Interspeech 2022
18-22 September 2022, Incheon, Korea

Use of Nods Less Synchronized with Turn-Taking and Prosody During Conversations in Adults with Autism

Keiko Ochi[1], *Nobutaka Ono*[2], *Keiho Owada*[3], *Miho Kuroda*[3], *Shigeki Sagayama*[4], *Hidenori Yamasue*[3.5]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2] Department of Computer Science, Tokyo Metropolitan University, Hino, Japan
[3] Faculty of Medicine, University of Tokyo, Tokyo, Japan,
[4] Professor Emeritus, University of Tokyo, Tokyo, Japan,
[5] Department of Psychiatry, Hamamatsu University School of Medicine, Hamamatsu, Japan
ochi.keiko.5f@kyoto-u.ac.jp, yamasue@hama-med.ac.jp

Abstract

Autism spectral disorder (ASD) is a highly prevalent neurodevelopmental disorder characterized by deficits in communication and social interaction. Head-nodding, a kind of visual backchannels, is used to co-construct the conversation and is crucial to smooth social interaction. In the present study, we quantitively analyze how head-nodding relates to speech turn-taking and prosodic change in Japanese conversation. The results showed that nodding was less frequently observed in ASD participants, especially around speakers' turn transitions, whereas it was notable just before and after turn-taking in individuals with typical development (TD). Analysis using 16 sec of long-time sliding segments revealed that synchronization between nod frequency and mean vocal intensity was higher in the TD group than in the ASD group. Classification by a support vector machine (SVM) using these proposed features achieved high performance with an accuracy of 91.1% and an F-measure of 0.942. In addition, the results indicated an optimal way of nodding according to turn-ending and emphasis, which could provide standard responses for reference or feedback in social skill training for people with ASD. Furthermore, the natural timing of nodding implied by the results can also be applied to developing interactive responses in humanoid robots or computer graphic (CG) agents.

Index Terms: dialogue, nod, prosody, turn-taking, autism spectrum disorders

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder, and a recent study reports its prevalence to be one in 54 children [1]. The core symptoms of ASD include deficits in social communication and interactions, including nonverbal communicative behaviors such as prosody. Many studies have attempted to quantify social interaction deficits [2]-[5] to provide objective measures for the symptoms; however, the most widely used assessment methods currently rely on subjective and qualitative rating by qualified experts.

Some studies have investigated the characteristics of ASD in speech, expecting its communication and interaction deficits to be presented [3]-[5]. A meta-analysis revealed the characteristics of ASD by the means and standard deviations (SDs) of fundamental frequency (F_0) and speech intensity [6]; however, previous studies comparing whole-session means and variances have reported inconsistent results. In conversation analyses, children with ASD tend to take a longer gap in turn-taking than those with typical development (TD) [7]. Individuals with ASD also show a lower acoustic/prosodic synchrony with interlocutors than those with TD [8]-[10]; children with TD align their speech rate more similarly than those with ASD [9].

In addition to speech, the poor production of gestures [11], facial expressions [12][13], eye gazes [14], and head-noddings are also known to characterize individuals with ASD. Eye gazes or facial expressions have been quantified using computer vision in many studies [15]-[18]. Head movement and nodding were analyzed in a study by Zhao *et al*., in which children with ASD were observed while answering questions, and high classification performance was achieved [19].

Repeated nods can be considered a specific means of social interaction regardless of ASD, which was shown in an investigation on adults [20]. In this study, frequent nodding appeared to occur at 'strong' boundaries of utterances; furthermore, its frequency increased with the social distance between the listener and the interlocutor as if the distance was compensated by a careful and respectful response [20]. In the same context, some studies attempted to detect nods to incorporate them into human-robot interaction [21].

Intrapersonal synchrony between head movement and speech has been shown in populations with TD in studies [22][23], that focused on the frame-by-frame alignment of these features on the basis of short time frames. However, the relationship between nods and speech in a longer period is still unclear.

Some studies focused on the quality or characteristics of head movement in terms of developmental disorders. Martin and colleagues conducted objective analyses on the head movement of children with and without ASD. They showed that children with ASD perform faster head-turning and inclination than peers with TD [24]. Zhao *et al*. used machine learning techniques to classify children with ASD and TD by employing a training model with some head-movement features [19]. By measuring the range of nodding/shaking and head rotation, they revealed that children with ASD move their heads in specific manners. However, nods during natural conversation have not been quantitively evaluated well.

In this study, focusing on turn-taking and intrapersonal synchrony, we quantitatively analyzed head-noddings and corresponding vocal features, comparing ASD and TD individuals in interactive conversation settings. We measured (1) the frequency of nods around turn-takings and (2) the synchrony between 16-sec blockwise frequencies of nods and the persons'/interlocutors' prosody. The difference in the characteristics between ASD and TD groups can be applied to social skill training (SST) for those with difficulties in communication. In addition, this qualification technique in social interaction can make it easier to perform objective assessment and evaluate efficacy in SST or other clinical treatments for ASD as well as raise the possibility of its automatic diagnosis.

.

10.21437/Interspeech.2022-11388

2. Participants and Data Acquisition

The video data was collected from the baseline of a clinical trial of intranasal oxytocin (i.e., the data collected before the start of administering oxytocin or placebo) in the University of Tokyo Hospital (UMIN000015264). Oxytocin is a kind of peptide hormone secreted via the pituitary gland that promotes milk ejection. Many oxytocin receptors are distributed in the brains of both men and women, and studies on healthy subjects have reported that oxytocin helps build trust with others. In this clinical trial, the effects of oxytocin on the core symptoms of ASD were detected.. The participants were informed beforehand through written informed consent. We recorded the conversations during semi-structured interviews included in the Autism Diagnostic Observation Schedule (ADOS). ADOS is one of the widely used gold-standard assessment tools for ASD. A psychiatrist (HY) administrated all activities related to ADOS on the participants.

The recorded participants were 65 male adults with ASD aged 18-48 years and 17 age-matched male controls aged 21-34 years. The intelligence quotient (IQ) level and socioeconomic status (SES) did not significantly differ between the two groups. They received Activity 7 in ADOS Module 4. This activity consists of questions and answers about emotions, e.g., "What do you feel angry about?" We excluded the data of one participant who withdrew his agreement and two whose recordings were incomplete. One participant, whose numbers of turn-takings and utterances were sufficient for analysis, was also excluded.

The participant sat behind a table with the ADOS administrator sitting on an adjacent side of the table (see Figure 1). A video camera recorded the two speakers from the front of the participant. Each person wore a wireless lavalier microphone around his collar to obtain an optimal signal-to-noise ratio. Each voice was simultaneously recorded to stereo channels of video in PCM WAV format.

3. Data Analyses

3.1. Video analysis

Let $y_i[k]$ be the y coordinate of the i-th landmark at the k-th video frame extracted using dlib library [25]. We first took the spatial mean of vertical coordinate values of the landmarks to reduce the effect of jaw movement from articulation and movement according to facial expression. $y[k]$ can be written as

$$y[k] = \frac{1}{N}\sum_{i=1}^{N} y_i[k] \quad , \tag{1}$$

where N is the number of landmarks per frame (=68). After that, we applied a 5-point temporal median filter for denoising. Let $\tilde{y}[k]$ be the resultant signal.

In this study, we regard the negative peaks of the time series of $y[k]$ as nods. The points where the first derivatives changed from negative to positive in a range wider than a threshold were detected as negative peaks. The Δ parameter of $\tilde{y}$ $[k]$ (the first derivative of the k-th frame) was calculated

3.2. Speech analyses

3.2.1. *Turn-taking timing*

We extracted turn transitions from the speech signals. Firstly, the boundaries of inter-pausal units (IPUs), speech segments divided by pauses longer than 200 msec, were manually annotated. We set the pause threshold as 200 msec not to separate the silence caused by geminate consonants in Japanese. Secondly, the starting and ending points of turns were automatically calculated, as shown in Figure 2. In this

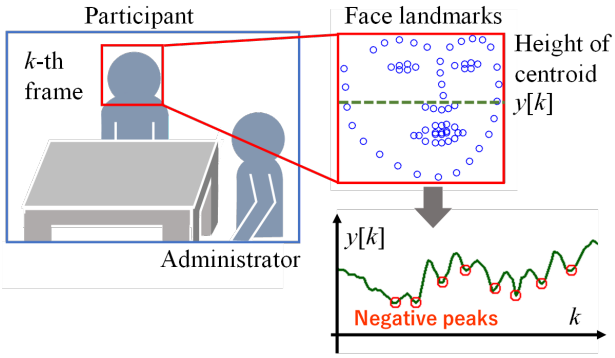


Figure 1: *Nod count from video recordings.*

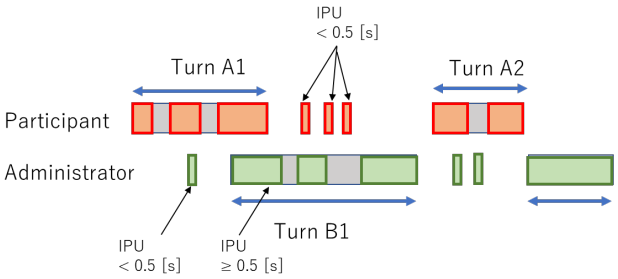


Figure 2: *Turn transition detection.*

study, we regarded a segment consisting of a series of IPUs as a turn, assuming that a turn is given when the produces an utterance longer than a threshold. The threshold was heuristically set to 0.5 sec. Thus, utterances shorter than 0.5 sec were regarded as backchannels.by subtracting the value of the posterior frame as

$$\Delta[k] = \tilde{y}[k+1] - \tilde{y}[k] \tag{2}$$

3.2.2. *Prosodic features*

We extracted log F_0 and intensity in dB for each analysis frame referring to Praat [26]. The frame step was 10 msec. We automatically set the F_0 ceiling and floor referring to [27]. After extracting F_0 with fixed floor and ceiling, the extraction was conducted again with an F_0 floor of 0.75 times the lower quantile and an F_0 ceiling of double the upper quantile. The pitch floor and ceiling were manually changed for two participants. Frames with manually detected overlapped speech were excluded from the subsequent analyses of the prosodic features.

3.3. Block analyses

To capture the long-term characteristics of prosody and nod frequency, we conducted a 'blockwise' analysis (see our previous study [1][2]). We defined a block as a 16-sec long segment and the blocks are half-overlapped as shown in Figure 3. The blockwise mean of the prosodic feature (intensity or log F_0 at the j-th block was calculated as

$$p_B[j] = \frac{1}{\#\Gamma_B[j]} \sum_{i \in \#\Gamma_B[j]}^{n_B} p[i] \quad , \tag{3}$$

where $\Gamma_B[j]$ represents the set of frame numbers included in the j-th block. n_B is the number of frames included in a block. The frames including multiple speakers' or unvoiced utterance were eliminated from $\Gamma_B[j]$. The total number of nods was also calculated for each block. The numbers were averaged within

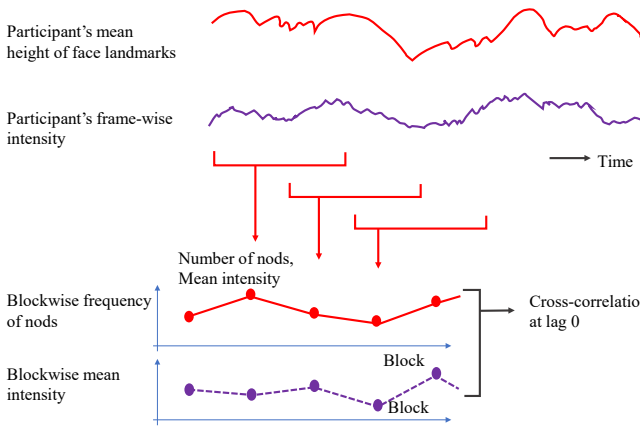


Figure 3: *Block analyses of nod frequency (the total number per block) and prosody*

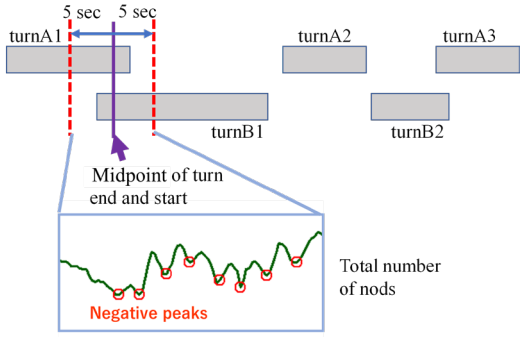
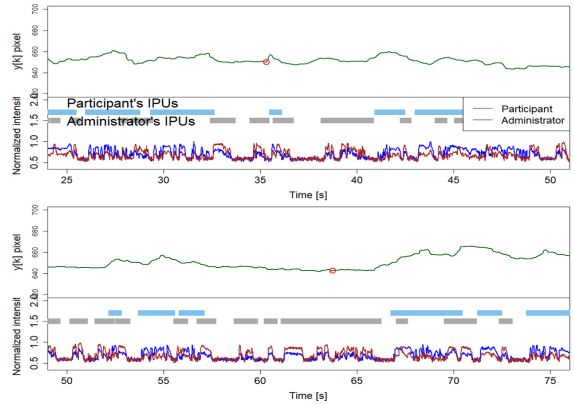
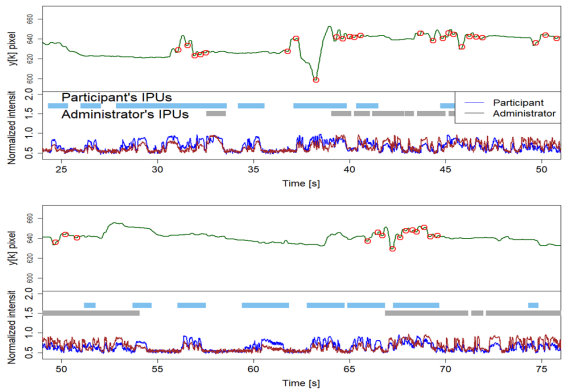


Figure 4: *Counting of nods around a turn transition*



(a) ASD



(b) TD

Figure 5: *Examples of mean height of the centroid of face landmarks of participants with (a) ASD and (b) TD, and intensity contours of two speakers. Red circles represent the detected nods.*

a session to obtain the mean frequency of nods as

$$\bar{f}_B = \frac{1}{N_B}\sum_{j=1}^{N_B} f_B[j] \ , \qquad (5)$$

where N_B and $f_B[j]$ represent the number of blocks within the session and the frequency of nods at the j-th block, respectively.

3.4. Relationship between nods and speech

3.4.1. *Frequency of nods around turn-taking*

In this study, we define the timing of the turn transition as the midpoint between the ending point of the last IPU of a turn and the starting point of the first IPU in the next turn. Figure 4 shows the calculation of the frequency of nods around turn transitions. We counted the nods within the segment from 5 sec before to 5 sec after the turn transition timing. For each participant, the mean of this number was calculated within a session.

3.4.2. *Cross-correlation with blockwise prosodic features*

To measure the synchrony between nod frequency and prosodic features, we calculated the cross-correlation at lag 0 defined as

$$r = \frac{\sum_{j=1}^{N_B}(p_B[j]-\overline{p_B})(f_B[j]-\bar{f}_B)}{\sqrt{\sum_{j=1}^{N_B}(p_B[j]-\overline{p_B})^2 \sum_{j=1}^{N_B}(f_B[j]-\bar{f}_B)^2}} \ , \qquad (3)$$

$$\overline{p_B} = \frac{1}{N_B}\sum_{j=1}^{N_B} p_B[j] \qquad (4)$$

3.5. Statistical analysis

We performed intergroup comparisons of the four proposed features between ASD and TD groups, using t-tests ($p < 0.05$). The p-values were adjusted using the Benjamini-Hochberg method [28] to control the type I errors in multiple comparisons. The effect size (Hedge's g) was obtained to evaluate how the distributions of the two groups overlap. We also conducted a discrimination analysis of the two groups using a support vector machine (SVM) with the proposed features. We used a linear kernel and weighting according to the data sizes of the two groups. We evaluated the SVM performance by leave-one-out cross-validation.

4. Results

Figure 5 showed the examples of the contour of the mean landmark height ($[k]$) of participants with ASD and TD. As can be seen in the lower panel, a participant with TD produced repeated nods around 32, 40, 45, 68 sec, which appeared to be coincident with turn-taking. In contrast, the participant with ASD showed no notable repeated nods.

Figure 6 shows examples of the blockwise mean of the intensity and frequency of nods. The contours of the two features are more synchronous for the participant with TD than for one with ASD; most negative peaks co-occur in the conversation of the participant with TD.

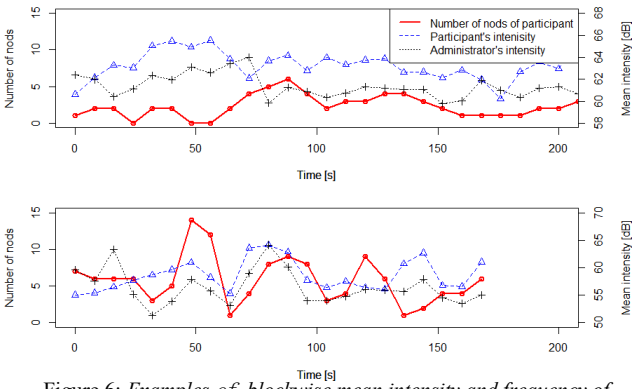


Figure 6: *Examples of blockwise mean intensity and frequency of nods of participants with ASD (top panel) and TD (bottom panel).*

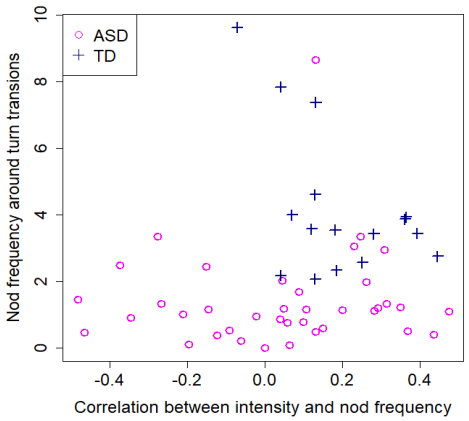


Figure 7: *Correlation between blockwise mean of intensity and nod frequency and nod frequency around turn transition.*

Table 1 shows the result of comparing the proposed features between the ASD and TD groups. Except for the correlation between the blockwise mean of log F_0 and the nod frequency, all features showed significant differences between the two groups. The mean frequency of nods within a session and around turn transitions was higher in people with TD than those with ASD. The effect size evaluated by Hedge's g was more than 1.0 in these features. The correlation between the blockwise mean of intensity and the nod frequency was also higher in the TD group than in the ASD group. The Hedge's g is slightly high in this feature.

Regarding the classification, the accuracy of using SVM was 91.1%, and the F-measure was 0.942, with the nod frequency around turn transition and the correlation between the blockwise mean of intensity and the nod frequency, showing the best performance among all possible combinations of features. Figure 7 shows the scatterplots of these two features. The participants with TD are distributed in a small area, whereas the distribution of the ASD group is widely spread in terms between the correlation of intensity and nod frequency.

5. Discussion

For the TD group, the number of head noddings per second was 0.45 around turn transitions, whereas the whole-session mean was 0.39. These results are consistent with the findings by Ishii *et al.* [20] indicating that adults with TD utilize nods, especially just before and after turn-taking and turn-giving in conversation, probably to show their understanding.

In the individuals with ASD, the frequency of head-nodding was almost five times lower than that in the TD individuals: 0.091 and 0.088 times per second at the turn transition and in a whole session, respectively. Their nod frequency was almost the same throughout the conversation, that is, not affected by turn-changing or turn-keeping in speech.

The correlation between blockwise mean of intensity and nod frequency was observed in the TD group (0.16). The speakers with TD align their nods to their voicing even in terms of long-time course. They may increase their frequency of head-nodding according to the degree of activation or emphasis of their talk. On the other hand, synchrony between nods and prosody only slightly appeared in the ASD group.

Since these three features mentioned above, but not the correlation between the log F_0 and nodding frequency, showed reasonably large effect sizes, they could distinctly reflect the difficulty of ASD individuals' difficulty in social interaction. These features obtained by monitoring head-nodding could be an effective measure to help ASD individuals with SST by giving instructions to, for example, add nods around turn transitions and at hot spots with high voice volume.

6. Conclusions

In the present study, we quantified the difference in the use of head-nodding between adults with ASD and TD in Japanese conversation. The use of head-nodding in other languages or cultures is required to be investigated. Future works also include utilizing the nodding strategy for the automatic diagnosis of ASD by combining it with other speech features. This quantification technique used to detect nodding can be applied to SST for its feedback and to contribute to facilitating effective interaction.

7. Acknowledgements

This research was supported by Japan Agency for Medical Research and Development (AMED) under Grant Number JP16dm0107134. This work was also partially supported by a JSPS KAKENHI Grant-in-Aid for Scientific Research (A) (Grant Number: 20H00613).

Table 1: *Result of intergroup comparison. Note that the numbers of nods around turn transitions were measured within a 10-sec segment, whereas other features are the numbers of times per 16 sec.*

Feature	Group mean		Adjusted	Hedge's
	ASD	TD	*p*-value	*g*
Mean frequency of nods	1.45	6.27	7.9×10^{-5}*	**1.89**
Nod frequency around turn transition	0.88	4.45	4.4×10^{-5}*	**2.18**
Correlation between blockwise mean of intensity and nod frequency	0.024	0.157	0.029*	0.67
Correlation between blockwise mean of log F_0 and nod frequency	-0.005	0.003	0.91	0.034

8. References

[1] M. J. Maenner, K. A. Shaw, J. Baio, A. Washington, M. Patrick, M. DiRienzo, *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring Network, 11 Sites, United States, 2016," *Morbidity and Mortality Weekly Report Surveillance Summaries,* vol. 69, no. 4, pp. 1–12J, 2020.

[2] T. Plötz, N. Y. Hammerla, A. Rozga, A. Reavis, N. Call, and G. D. Abowd, "Automatic assessment of problem behavior in individuals with developmental disabilities," *Proc. 2012 ACM Conference on Ubiquitous Computing*, pp. 391–400, 2012.

[3] D. Bone, J. Mertens, E. Zane, S. Lee, S. S. Narayanan, and R. B. Grossman, "Acoustic-prosodic and physiological response to stressful interactions in children with autism spectrum disorder," *Proc. INTERSPEECH*, pp. 147–151, 2017.

[4] S. Cho, M. Liberman, N. Ryant, M. Cola, R. T. Schultz, and J. Parish-Morris, "Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations," *Proc. INTERSPEECH 2019,* pp. 2513–2517, 2019.

[5] H. Drimalla, T. Scheffer, N. Landwehr, L. I. Baskow, S. Roepke, B. Behnia, and I. Dziobek, "Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (SIT)," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[6] 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 Research*, vol. 7, no. 3, pp. 384–407, 2017.

[7] P. Heeman, R. Lunsford, E. Selfridge, L Black, L. van Santen, "Autism and interactional aspects of dialogue," *Proc. 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* pp. 249–252, 2010.

[8] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic entrainment in conversations of verbal children and teens on the autism spectrum," *Frontiers in Psychology*, vol. 11, 582221, 2020.

[9] C. J. Wynn, S. A. Borrie, and T. P. Sellers, "Speech rate entrainment in children and adults with and without autism spectrum disorder," *American Journal of Speech-Language Pathology*, vol. 27, no. 3, pp. 965–974, 2018.

[10] K. Ochi, N. Ono, K. Owada, M. Kojima, M. Kuroda, S. Sagayama, and H. Yamasue, "Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder," *PloS One*, vol. 14, no. 12, e0225377, 2019.

[11] A. Mishra, V. Ceballos, K. Himmelwright, S. McCabe, and L. Scott, "Gesture production in toddlers with autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 51, no. 5, pp. 1658–1667, 2021.

[12] D. A. Trevisan, M. Hoskyn, E. Birmingham, "Facial expression production in autism: A meta‐analysis," *Autism Research,* vol. 11, no. 12, pp. 1586–1601, 2018.

[13] K. Owada, T. Okada, T. Munesue, and M. Kuroda, T. Fujioka, *et al.*, "Quantitative facial expression analysis revealed the efficacy and time course of oxytocin in autism," *Brain,* vol. 142, no. 7, pp. 2127–2136, 2019.

[14] T. Ruffman, W. Garnham, and P. Rideout, "Social understanding in autism: Eye gaze as a measure of core insights," *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 42, no. 8, pp. 1083–1094, 2001.

[15] J. Hashemi, M. Tepper, T. Vallin Spina, A. Esler, V. Morellas, N. Papanikolopoulos, *et al.*, "Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants," *Autism Research and Treatment*, 935686. 2014.

[16] M. Del Coco, M. Leo, P. Carcagni, P. Spagnolo, P. Luigi Mazzeo, M. Bernava, *et al.*, "A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders," *Proc. IEEE International Conference on Computer Vision Workshops*, pp. 1401–1407, 2017.

[17] K. Campbell, K. L. Carpenter, J. Hashemi, S. Espinosa, *et al.*, "Computer vision analysis captures atypical attention in toddlers with autism," *Autism*, vol. 23, no. 3, pp. 619–628, 2019.

[18] G. Dawson, K. Campbell, J. Hashemi, S. J. Lippmann, V. Smith, K. Carpenter, *et al.*, "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," *Scientific Reports*, vol. 8, no. 1, pp. 1–7, 2019.

[19] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, *et al.*, "Identifying Autism with Head Movement Features by Implementing Machine Learning Algorithms," *Journal of Autism and Developmental Disorders 2021*, pp. 1–12, 2011.

[20] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of relationship between head motion events and speech in dialogue conversations," *Speech Communication,* vol. 57, pp. 233–243, 2014.

[21] E. Wall, L. Schillingmann, and F. Kummert, "Online nod detection in human-robot interaction," *Proc. 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 811–817, 2017.

[22] U. Hadar, T. J. Steiner, F. and Clifford Rose, "Head movement during listening turns in conversation," *Journal of Nonverbal Behavior*, vol. 9, no. 4, pp. 214–228, 1985.

[23] B. Xiao, P. G. Georgiou, C. C. Lee, B. Baucom, and S. S. Narayanan, "Head motion synchrony and its correlation to affectivity in dyadic interactions," *Proc. 2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2013.

[24] K. B, Martin, Z. Hammal, G. Ren, J. F. Cohn, J. Cassell, M. Ogihara, *et al*., "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular autism,* vol. 9, no. 1, pp. 1–10, 2018.

[25] Face landmark detection tool Dlib http://dlib.net

[26] P. Boersma "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound," *Proceedings of the Institute of Phonetic Sciences*. vol. 17, no. 1993, pp. 97–110, 1993.

[27] C. De Looze and D. J. Hirst "Detecting Changes in Key and Range for the Automatic Modelling and Coding of Intonation," *Proc. Speech Prosody 2008,* 2008.

[28] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 57, no. 1. pp. 289–300, 1995.