Interspeech 2024
1-5 September 2024, Kos, Greece

Acoustic changes in speech prosody produced by children with autism after robot-assisted speech training

Sarah Si Chen[17], Bruce Xiao Wang[27], Yitian Hong[1], Fang Zhou[1], Angel Chan[1], Po-yi Tang[1], Bin Li[3], Chen Chunyi[4], James Cheung[4], Yan Liu[5], Zhuoming Chen[6]

[1]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
[2]Department of English and Communication, Hong Kong Polytechnic University
[3]Department of Linguistics and Translation, City University of Hong Kong
[4]Department of Biomedical Engineering, Hong Kong Polytechnic University
[5]Department of Computing, Hong Kong Polytechnic University
[6]Brain Hospital, The First Hospital of Jinan University
[7]Research Institute for Smart Ageing, The Hong Kong Polytechnic University
{sarah.chen/brucex.wang/angel.ws.chan/chunyi.wen/james.chungwai.cheung/yan.liu}@polyu.edu.hk; {ytian.hong/fang-vf.zhou/po-yi.tang}@connect.polyu.hk; binli2@cityu.edu.hk; zm120tchzm@qq.com

Abstract

Children with Autism exhibit distinct speech prosody, perceived as monotone and they are reported to show deficits in focus marking. The current study designed a robot-assisted training with controlled social interactions aiming to enhance the prosody of children with Autism speaking a tonal language, Cantonese, specifically on focus marking. 20 autistic and 23 typically-developing (TD) children participated in this study. Only the autistic group received training. Stimuli were designed for training, pre- and post-training production. Acoustics of target words were extracted and analysed using linear mixed-effects models examining effects of training and clinical status. Children with Autism improved in signalling sentence prominence using duration but not f0 and intensity. Variability suggests that certain acoustic cues are more challenging. Comparing to TD children's focus marking patterns, autistic children's variability may also stem from their ongoing prosodic profile development.

Index Terms: speech prosody, robot-assisted training, autism spectrum disorder

1. Introduction

1.1. Focus marking

Focus marking, one of the important prosodic functions, is often used to signal the prominence of information in an utterance [1]. and can be categorised into broad focus (i.e., all parts of the utterance have equal importance), narrow focus (i.e., parts of the utterance have higher importance) and contrastive focus (i.e., when new information is introduced) [2]. Table 1 gives an example of focus marking as a function of questions being asked and on-focus words are in bold font. Depending on the questions, narrow and contrastive focus can be further divided into initial (Table 1. b1 and c1), medial (Table 1. b2 and c2) and final focus (Table 1. b1 and c1).

Numerous studies have explored the acoustic correlates of focus marking among the typically developing (TD) population, e.g., f0 [3], intensity [4], duration [5]. In general, syllables at on-focus position have higher f0 (e.g., on-focus f0 expansion; OFE [1]) and intensity and longer duration than those at off-focus positions (e.g., post-focus f0 compression; PFC). Some studies further suggest that the variation in focus marking depends not only on whether the syllable is on-focus or off-focus, but also on different focus types (e.g., Table 1), with contrastive focus having the highest f0, intensity, and duration, followed by narrow focus and broad focus [3], [6], [7].

1.2. Autism spectrum disorder

Autism spectrum disorder (ASD) is an early-appearing neurodevelopmental disorder that affects various aspects of life [8], [9]. The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders [10] stated that deficit of social communication and interaction is one of the major symptoms of ASD, and ASD individuals might have difficulty in understanding social norms, leading to failure of carrying out normal verbal communication. The main difficulties in understanding emotion norms in communication for ASD individuals resulted from the lack of ability to perceive and produce reciprocal prosodic cues (e.g., focus marking). Prosody portrays the suprasegmental features of speech and serves an important role in communicative functions, e.g., affective, pragmatic and syntactic [11]. Change in the prosody would lead to change in the meaning of the same utterance [2].

Children with Autism often have different prosodic patterns comparing to their TD counterparts [12], and among studies comparing prosody between the autistic and the TD population, f0, intensity and duration are the most common acoustic features explored [8]. [13] employed tasks to investigate autistic children's ability of using prosody to indicate the prominence of information in an utterance. In the *Focus* task, essentially a task designed to test autistic children's ability of using contrastive focus, they found that autistic children had higher mean f0, f0 range, intensity as well as longer utterance duration than the TD children; however, the differences were not statistically significant. On the other hand, [14] showed that the autistic group had a significant higher mean f0 and f0 range as well as longer word duration than the TD group in their study. Other studies, although the main purposes were not to investigate the differences in acoustic measures between the

10.21437/Interspeech.2024-68

autistic and the TD children, have also reported similar patterns [12], [15], [16] [17].

Table 1: *Examples of prompt questions and target sentences in relation to different focus conditions.*[1]

Focus Condition	Prompt Questions	Target Sentence
a. Broad	What do you see in the picture?	Mr. Cheung is flying the plane.
b. 1) Initial narrow	Who is flying the plane?	**Mr. Cheung** is flying the plane.
b. 2) Medial narrow	What is Mr. Cheung doing to the plane?	Mr. Cheung is **flying** the plane.
b. 3) Final narrow	What is Mr. Cheung flying?	Mr. Cheung is flying the **plane**.
c. 1) Initial contrastive	Miss Chan is flying the plane?	**Mr. Cheung** is flying the plane.
c. 2) Medial contrastive	Mr. Cheung is buying the plane?	Mr. Cheung is **flying** the plane.
c. 3) Final contrastive	Mr. Cheung is driving the bus?	Mr. Cheung is **flying** the plane.

1.3. Robot-assisted training with social interactions Focus Marking and acoustic correlates

Social interaction is essential to language acquisition [18] , and previous studies showed that training with simulated social interactions in language acquisition leads to positive behavioural outcome and brain changes [19], [20]. Nevertheless, little is known about the effectiveness of social interaction on the autistic population, especially when targeting a specific aspect of language acquisition, namely prosodic focus marking. The current study employed a robot-assisted training with simulated social interactions aiming to investigate the effectiveness in training speech prosody produced by autistic children.

Autistic children have difficulties in processing and integrating multisensory information such as facial expressions and gestures during social interaction [21], [22] and there is no empirical study showing if autistic children would benefit from and whether they are ready to integrate such multisensory information. Further, human behaviour (e.g., facial expression) is not entirely predictable when it comes to communication and social interaction [23], [24]. Therefore, using a robot enables us to control the consistency of such behavioural information and focus only on the effect of social interaction on the production of speech prosody among autistic children. Figure 1 shows the *Furhat* robot used in the current study.

Figure 1: *Example of Furhat setup used for training where the child and Furhat sit face-to-face.*

Given that speech prosody plays a crucial role in communication, speech training of this sort conducted in the current study is important for the development of prosodic skills among autistic children.

2. Method

2.1. Participants

The current study recruited 20 and 23 Cantonese-speaking autistic and TD children respectively. Participants were assessed using the Autism Spectrum Quotient (AQ): Children's Version [25] and non-verbal Intelligence Quotient (IQ). Table 2 gives the mean and standard deviation of age, IQ and AQ scores.

2.2. Production and stimuli

Fifteen target sentences and prompt questions were designed to elicit the desired types of focus (i.e., broad, narrow, contrastive). Each sentence describes an on-going action with corresponding pictures depicting the content of the target sentences. Six of them were used for training, while the whole fifteen sentences were employed for pre- and post-training production. During the production sessions, the participants were instructed to answer questions from the robot using the target sentences and corresponding pictures. For example, a partially blocked picture (Figure 2. left panel) was firstly shown to participants with robot asking, "邊個揸飛機 who is flying the plane?". A complete picture (Figure 2. right panel) was then given and participants were expected to answer "[張生]$_{focus}$ 揸飛機 [Mr. Cheung]$_{focus}$ is flying the plane." with focus on the initial subject. Robot's production was pre-recorded by a native Hong Kong Cantonese speaker and triggered by the experimenter manually. Each participant had 210 stimuli (15 target sentences * 7 focus types * 2 repetitions). The production test was recorded once for the TD group, while the autistic group was recorded before and after the training.

Table 2: *Mean and standard deviation of age, IQ and AQ scores of participants.*

	TD	ASD
Age	7.82	9.53
	±1.16	±1.61
IQ	113.61	109.89
	±13.82	±18.91
AQ	58.95	80.83
	±14.64	±28.44

Figure 2: *An example eliciting narrow focus in the production test.*

2.3. Procedure

Both production and training were conducted in a sound-proof booth at the speech lab of the Hong Kong Polytechnic University. Audio Technica AT2035 condenser microphone and Steinberg UR22mkII USB Audio Interface were used to record participants' speech in the two production tasks with the sample rate of 44100 Hz in Audacity.

There were two training sessions, each comprising four blocks, with a minimum interval of at least 24 hours between them. Different stimuli sentences were utilised in the two training sessions while the procedure of training remained identical. During the first two blocks, participants were instructed to

[1] Note that these are translated version of the stimuli, the original stimuli are in Hong Kong Cantonese.

prompt the robot based on questions from the production stimuli and to identify the focus type (e.g., narrow pre-focus) based on the robot's responses. Participants were instructed to click on the corresponding button for the identified focus (Figure 3.). In the subsequent two blocks, participants continued to prompt the robot using questions from the production stimuli. However, their task shifted to determining whether the robot provided the correct focus type based on its response, and the feedback was provided by the robot after the child's judgement. The robot consistently delivered congruous answers (e.g., narrow focus prompt with narrow focus prosody answer) in the first two blocks. In contrast, during the second two blocks, the robot could offer both congruous and incongruous responses (e.g., narrow focus prompt with broad focus prosody answer).

Figure 3: *Example of the interface used for lab perceptual training.*

2.4. Segmentation, feature extraction and statistical analyses

Production test was recorded and saved as *wav* files and target words were manually segmented using a TextGrid in Praat [26]. Word duration, mean f0 and intensity were extracted using ProsodyPro [27] with f0 range from 75 Hz to 600 Hz.

Linear Mixed-Effects Models (LMM) were fitted to acoustic data and a Likelihood Ratio (LR) test was conducted to investigate the significance of explanatory variables on the response variables. Word duration, mean f0, mean intensity and f0 range were assigned as the response variables in each model, while groups (i.e., ASD training, ASD control and TD) and focus conditions (i.e., narrow/contrastive pre-, on- and post-focus) were assigned as explanatory variables. Model fitting started with null intercept but two random effects (i.e., participants and words), and gradually adding each explanatory variable to the model followed by LR tests to investigate whether there is a significant difference between models with and without certain fixed effects. A post-hoc comparison was followed if any significant effect was reached. The optimal model was chosen based on the lowest Akaike information criterion [24] and significant p value. The implementation of LMM fitting, LR test and post-hoc comparison were carried out using lmer4[28], anova[29] and emmeans[30] in R [28]. The LMM analyses were conducted for both pre- and post-training sessions for the ASD training and control groups.

3. Results

Figures 3 – 5 show the mean differences in acoustic values between broad focus and the other focus conditions (i.e., narrow and contrastive on- and post-focus) for three groups. Negative numbers indicate a higher value for words under broad focus conditions, and the asterisks indicate the degree of significance (i.e., '***' 0.001, '**' 0.01, '*' 0.05). Only the comparisons between broad focus and on- and post-focus are given in the plots.

3.1. Duration

For the TD group, LMM analyses showed that the main effect of focus condition had a significant effect on mean duration (χ^2 = 335.85; df = 6; p < 2.2e-16). A *post-hoc* comparison showed that the TD group produced post-focus words with significantly shorter duration than the broad focus words under both narrow and contrastive conditions, whereas they only produced contrastive on-focus words significantly longer than broad focus words. Meanwhile, focus condition had a significant effect on mean duration for the autistic group in both pre- (χ^2 = 120.96; df = 6; p < 2.2e-16) and post-training (χ^2 = 208.78; df = 6; p < 2.2e-16) production. A *post-hoc* comparison revealed that the autistic group had significantly shorter post-focus words than broad focus words in pre-training production, while in post-training, they produced on-focus words with significantly longer duration than broad focus words under contrastive conditions.

3.2. F0

Among the TD children, LMM analyses showed that the main effect of focus condition had a significant effect on mean f0 (χ^2 = 1558.2; df = 6; p < 2.2e-16). A *post-hoc* comparison revealed that the TD group produced post-focus words with significantly lower mean f0 than broad focus words under both narrow and contrastive focus conditions and on-focus words with significantly higher mean f0 under contrastive focus conditions. For the autistic groups, the main effect of focus condition also had a significant effect on mean f0 in pre- (χ^2 = 126502; df = 6; p < 2.2e-16) and post-training (χ^2 = 703.44; df = 6; p < 2.2e-16) production. A *post-hoc* comparison suggested that the autistic group produced on-focus words with significantly higher mean f0 than broad focus words in pre-training production, while in post-training, they produced post-focus words with significantly lower mean f0 than broad focus words.

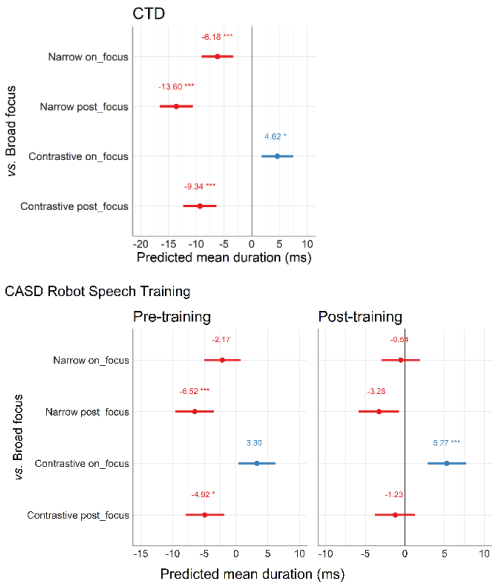

Figure 3: *Difference in predicted mean duration between broad focus and narrow and contrastive on- and post-focus across groups and training sessions.*

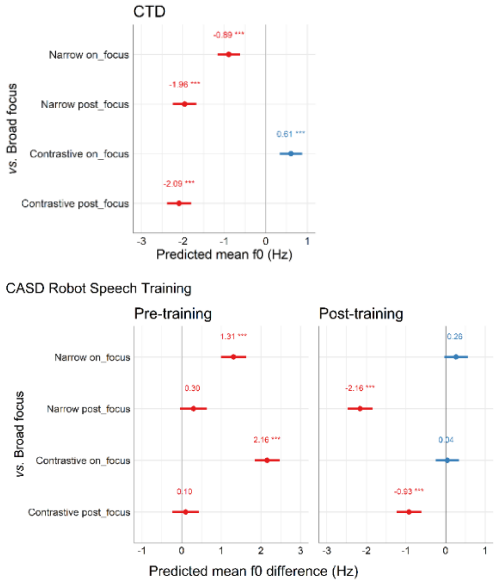

Figure 4: *Difference in predicted mean f0 between broad focus and narrow and contrastive on- and post-focus across groups and training sessions.*

3.3. Intensity

For the TD group, LMM analyses showed that the main effect of focus condition had a significant effect on mean intensity (χ^2 = 1377.8; df = 6; p < 2.2e-16), and a *post-hoc* comparison revealed that the TD children produced post-focus words with significantly lower mean intensity than broad focus words under narrow and contrastive conditions. Among the autistic group, focus condition had a significant effect on mean intensity in both pre- (χ^2 = 535.86; df = 6; p < 2.2e-16) and post-training (χ^2 = 387.94; df = 6; p < 2.2e-16) sessions. A *post-hoc* comparison revealed, in pre-training production, that the autistic group produced on-focus words with significantly higher and post-focus words with significantly lower mean intensity than broad focus words under both narrow and contrastive conditions.

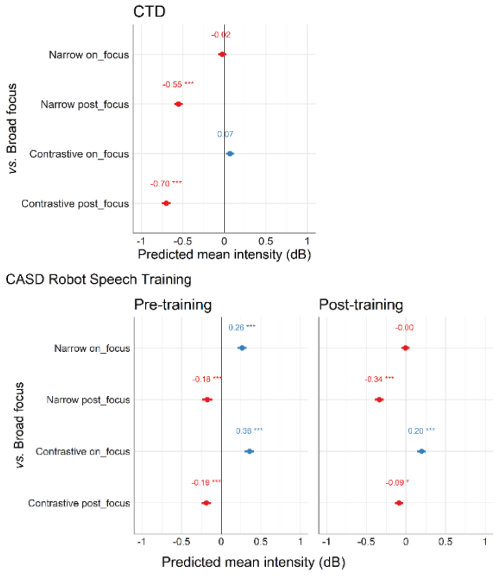

Figure 5: *Difference in predicted mean intensity between broad focus and narrow and contrastive on- and post-focus across groups and training sessions.*

4. Discussion

This section gives the descriptive of changes in the use of prosodic focus marking among the autistic children in relation to training sessions as well as the TD group and discusses the effect of training on ASD group's acquisition of prosodic focus marking.

In using duration to signal sentence prominence, the autistic children did not show longer duration to indicate on-focus words in pre-training sessions, but produced on-focus words with significantly longer duration than the broad focus words under contrastive focus condition (Figure 3, lower panel). This pattern in post-training sessions is more similar to that of the TD group where the on-focus words produced by the TD children had significantly longer duration than that of the broad focus words (Figure 3. upper panel). This alignment with the TD children might suggest a positive impact of the robot-assisted training on the use of duration in signalling sentence prominence for the autistic children. However, the TD children produced post-focus words with significantly shorter duration than the broad focus words, while this pattern is absent among the autistic children in the post-training session. The coordination of on-focus expansion along with post-focus compression necessitates the ability to oversee sentence production globally. Given that the ASD population has been reported to exhibit a bias toward local processing rather than global processing [31], manipulating both the on-focus target and post-focus element simultaneously might be challenging for them. The robot training successfully shifted their attention to the target word, albeit at the expense of manipulating the post-focus element. Future studies should consider training designs from a global perspective to enhance their ability to manage sentence production globally.

While there seem to be some improvement in the use of duration to signal sentence prominence among autistic children after robot-assisted training, the use of f0 and intensity appears to be more challenging for the autistic group to acquire. For example, the autistic children produced on-focus words with significantly higher f0 than the broad focus words in pre-training session, but not in post-training (Figure 4. Lower panel). In the use of intensity, the autistic children produced post-focus words with significantly lower intensity than the broad focus words in both pre- and post-training sessions. This pattern of lower intensity also resembles the TD group. Surprisingly, the autistic children produced on-focus words with significantly higher intensity than broad-focus words under both narrow and contrastive focus conditions in pre-training session. However, this pattern is only observed under contrastive focus condition in post-training session (Figure 5, Lower panel). This suggests that robot-assisted training did not have much positive impact on the autistic children's use of intensity for prosodic focus marking.

Overall, it seems that the prosodic focus marking produced by the autistic children does not consistently align with the desired direction following the robot-assisted training. This inconsistency among the autistic cohort is likely to suggest that certain acoustic cues pose greater challenges for acquisition compared to the others. For instance, the utilisation of duration to indicate sentence prominence might be more intuitively grasped than the use of f0 and intensity. By taking the focus

marking patterns observed in the TD children into consideration, this inconsistency within the autistic children can also be attributed to the ongoing developmental trajectory of their prosodic profile.

5. Conclusions

The current study investigated the effectiveness of using robot-assisted TSSI in improving speech prosody produced by Cantonese-speaking children with autism, showing improvement in using certain acoustic cues, such as duration, for indicating sentence prominence. Future research is needed to design stimuli and training programme focusing the training of specific acoustic cues, such as f0 and intensity, in speech prosody production.

6. Acknowledgements

This work was supported by Department of Chinese and Bilingual Studies at Hong Kong Polytechnic University [1-ZVRT; 1-ZE0D; 1-W08C], and partly supported by the SCOLAR, Education Bureau, HKSAR [K-ZB2P], RGC [A-PB1B] and Sin Wai Kin Foundation Limited (ZH5Z).

7. References

[1] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *J. Phon.*, vol. 33, no. 2, pp. 159–197, Apr. 2005, doi: 10.1016/j.wocn.2004.11.001.

[2] A. Cruttenden, *Intonation*. Cambridge University Press, 1997.

[3] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 1044–1098, Sep. 2010,

[4] J. Harrington, J. Fletcher, and M. E. Beckman, "Manner and place conflicts in the articulation of accent in Australian English," in *Papers in Laboratory Phonology*, vol. 5, Cambridge University Press, 2000, pp. 40–55.

[5] P. Lieberman, "Some Acoustic Correlates of Word Stress in American English.," *The Journal of the Acoustical Society of America*, vol. 32, no. 4, p. 451, 1960.

[6] D. Mücke and M. Grice, "The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation?," *J. Phon.*, vol. 44, pp. 47–61, May 2014.

[7] S. Roessig, B. Winter, and D. Mücke, "Tracing the Phonetic Space of Prosodic Focus Marking," *Front. Artif. Intell.*

[8] S. Z. Asghari, S. Farashi, S. Bashirian, and E. Jenabi, "Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Nov. 2021.

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

[10] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorders (DSM-5) 5th ed," Washington DC: American Psychiatric Association Publishing, 2013.

[11] S. Peppé, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and Expressive Prosodic Ability in Children With High-Functioning Autism," *J. Speech Lang. Hear. Res.*, vol. 50, no. 4, pp. 1015–1028, Aug. 2007.

[12] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum," *Front. Psychol.*, vol. 11, 2020, Accessed: Nov. 02, 2022.

[13] J. J. Diehl and R. Paul, "Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders," *Appl. Psycholinguist.*, vol. 34, no. 1, pp. 135–161.

[14] M. G. Filipe, S. Frota, S. L. Castro, and S. G. Vicente, "Atypical Prosody in Asperger Syndrome: Perceptual and Acoustic Measurements," *J. Autism Dev. Disord.*, vol. 44, pp. 1972.

[15] Y. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal Speech Spectrum and Increased Pitch Variability in Young Autistic Children," *Front. Hum. Neurosci.*, vol. 4, 2011.

[16] A. Nadig and H. Shaw, "Acoustic and Perceptual Measurement of Expressive Prosody in High-Functioning Autism: Increased Pitch Range and What it Means to Listeners," *J. Autism Dev. Disord.*, vol. 42, no. 4, pp. 499–511, Apr. 2012.

[17] R. B. Grossman, R. H. Bemis, S. D. Plesa, and -Flusberg Helen Tager, "Lexical and Affective Prosody in Children With High-Functioning Autism," *J. Speech Lang. Hear. Res.*, vol. 53, no. 3, pp. 778–793, Jun. 2010.

[18] P. Li and H. Jeong, "The social brain of language: grounding second language learning in social interaction," *Npj Sci. Learn.*, vol. 5, no. 1, Art. no. 1, Jun. 2020.

[19] H. Jeong *et al.*, "Learning second language vocabulary: neural dissociation of situation-based learning and text-based learning," *NeuroImage*, vol. 50, no. 2, pp. 802–809, Apr. 2010.

[20] P. K. Kuhl, F.-M. Tsao, and H.-M. Liu, "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning," *Proc. Natl. Acad. Sci.*, vol. 100, no. 15, pp. 9096–9101, Jul. 2003.

[21] S. Wallace, M. Coleman, and A. Bailey, "An investigation of basic facial expression recognition in autism spectrum disorders," *Cogn. Emot.*, vol. 22, pp. 1353–1380, Nov. 2008.

[22] A. L. Hubbard, K. McNealy, A. A. Scott-Van Zeeland, D. E. Callan, S. Y. Bookheimer, and M. Dapretto, "Altered integration of speech and gesture in children with autism spectrum disorders," *Brain Behav.*, vol. 2, pp. 606–619, 2012.

[23] A. M. Alcorn *et al.*, "Educators' Views on Using Humanoid Robots With Autistic Learners in Special Education Settings in England," *Front. Robot. AI*, vol. 6, 2019.

[24] A. Taheri, A. Meghdari, M. Alemi, and H. Pouretemad, "Human–Robot Interaction in Autism Treatment: A Case Study on Three Pairs of Autistic Children as Twins, Siblings, and Classmates," *Int. J. Soc. Robot.*, vol. 10, pp. 93–113, Jan. 2018.

[25] B. Auyeung, S. Baron-Cohen, S. Wheelwright, and C. Allison, "The autism spectrum quotient: Children's version ((R.W.S. Chan, W.S. Liu, K.K. Chung, C.S. Sheh & E.K.F. Woo Trans.)," *Journal of autism and developmental disorders*, vol. 38, no. 7, pp. 1230–1240, 2008.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2017.

[27] Y. Xu, "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis," in *Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France, 2013, pp. 7–10.

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, 2015.

[29] E. Arnhold, "Pakcage in the R environment for analysis of variance and complementary analyses," *Brazilian Journal of Veterinary Research and Animal Science*, no. 50, 2013.

[30] R. Lenth, "_emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.8.1-1." 2022.

[31] F. Happé and U. Frith, "The weak coherence account: detail-focused cognitive style in autism spectrum disorders," *J. Autism Dev. Disord.*, vol. 36, no. 1, pp. 5–25, Jan. 2006.