Pitch and Volume Stability in the Communicative Response of Adults with Autism

Keiko Ochi[1], Masaki Kojima[2], Keiho Owada[2], Nobutaka Ono[3], Shigeki Sagayama[2], and Hidenori Yamasue[4]
[1] Kyoto University, Kyoto, Japan, E-mail: ochi.keiko.5f@kyoto-u.ac.jp Tel: +81-75-753-7531
[2] University of Tokyo, Japan, Email: {keihou.ohwada@gmail.com, sagayama@ieee.org}, Tel: +81-3-3812-2111
[3] Tokyo Metropolitan University, Japan, E-mail: onono@tmu.ac.jp, Tel: +81-42-585-8418
[4]Hamamatsu University School of Medicine E-mail: yamasue@hama-med.ac.jp Tel: +86-53-435-2995

Abstract— **In this study, we investigate the characteristics of prosody in a newly developed speech experiment to objectively and quantitatively characterize prosodic features of autism spectral disorder (ASD). In the experiment, male adults with high-functioning ASD and age-, intelligence-level-matched males with typical development (TD) read aloud 29 brief scripts as if they were responding to the preceding auditory stimuli. On the basis of the hypothesis that autistic speakers tend to react in a stereotypical manner regardless of the situation, we quantitively evaluated the difference between their responses to stimuli of the same script by different speakers in terms of the prosodic features, such as fundamental frequency, intensity, and mora duration. The results showed that the individuals with ASD had similar pitch registers or volume levels, whereas those with TD reacted in higher maximum pitches or intensities to the male voice in some tasks. This result may be related to the recently reported stability of pitch control in autistic speakers. In contrast, regarding the 'disgust' emotion, the mean absolute error of intensity between the reactions to two actors was higher in the ASD group than in the TD group.**

I. INTRODUCTION

Autism spectrum disorder (ASD) is one of the widely prevalent neurodevelopmental disorders with a prevalence rate of 1 in 54 [1]. There are currently no approved medications for the core symptoms, including deficits in social communication and interactions, which sometimes cause secondary disorders and deterioration of the quality of life. The diagnosis and assessment of ASD require long hours of interviews with an experienced clinician (e.g., Autism Diagnostic Interview Revised (ADIR)). Although structured diagnostic tools are available to standardize the assessment, they are exhausting for the testees, depend on expertized subjective rating, and are not designed for repeated use with the assumption of deficits being stable.

Many studies point out the social communication characteristics specific for individuals with ASD [2]-[7]. Vocal paralinguistic features have been analyzed with other nonverbal information such as facial expressions and gestures because they are closely related to social communication and interaction. Determining such vocal features could be promising for developing a simple and easy-to-use method for objective, quantitative, and reproducible assessments with improved accuracy of diagnosis. The high assessment accuracy can contribute to further development of novel therapies by fine-grained evaluation of time-course changes in the severity of ASD core symptoms.

Recently, improved machine learning techniques have been successfully applied to the automatic diagnosis or assessment of the severity of ASD. The convolutional neural networks (CNNs) provide good performance for the estimation of severity of young children with ASD [7]. The Bidirectional Encoder Representations from Transformers (BERT) based vocabulary features can also be effective predictors used with audio features [8]. Levy *et al.* introduced the subsets of biobehavioral features effective for classification [9]. A semi-supervised classification method is also applied to overcome the limitation of datasets [9].

Many of past studies attempted to diagnose and assess ASD using acoustic and prosodic features of speech signals during word naming, narrative, spontaneous speech, or dialog [2]-[12]. Some studies use speech recordings of semistructured interviews in Autism Diagnostic Observation Schedule (ADOS) to control speech topics [7][10]-[13], which allow the testees to speak in their own words. However, a more strictly structured setting should be required to eliminate the variation of sentences that make the comparison of the prosody difficult.

For this reason, in our study, we developed a prototype of a set of brief scripts as an assessment tool to optimize quantitative and objective analyses of speech features specifically for adults with ASD. By making testees to read aloud predetermined scripts as a response to auditory stimuli in various situations, we can obtain highly structured speech data in terms of spoken text. By comparing the paralinguistic features in a certain text for one speaker and among the speakers, we can focus on their acoustic and prosodic control in specific situations and contexts.

Moreover, by the method provided here, we simplified and structured the test environment so that we were able to eliminate the variability in the performance of interviews. Specifically, our method does not require the testee to be interviewed by an interlocutor while they respond to consistent auditory stimuli on loudspeakers, which is in contrast with previous studies in which the interviewer had to interfere with the testee in order to elicit their social response.

The prosodic and acoustic features in children with ASD are more stable than those in their peers [14], which may be related to the asynchrony between autistic speakers and their

interlocutor of dialogs [13]. Thus, when speaking the same script, speakers with ASD are assumed to produce stable prosodic patterns. On the basis of this hypothesis, we measure the differences in prosodic features between the same-sentence responses to different vocal stimuli. We carry out classification analysis using a machine learning technique selecting the optimized set of prosodic features.

In Section II, we describe the experimental settings and recorded scripts. In Section III, we provide the methodology of speech feature extraction, significance tests between the ASD and TD groups, and classification analysis. We show the results of analyses in Section IV. Finally, we discuss and conclude this study.

II. EXPERIMENTS

A. Participants

Twenty Japanese males with ASD and aged 19-38 years and 22 male controls with typical development (TD) and aged 16-53 years participated in the experiment. A psychiatrist (HY) experienced in developmental disorders made the ASD diagnosis utilizing gold standards such as ADIR and ADOS. The participants with a history of TD were recruited with their intelligence level matching those of the ASD group and age and screened by a trained psychiatrist (M.K.) on the basis of the following exclusion criteria: presence and/or history of neuropsychiatric disorders. The Ethical Committee of the University of Tokyo Hospital approved this study (10245). After a complete explanation of the study, participants' mental capacity to consent was confirmed by a psychiatrist (M.K.), and written informed consent was obtained from all participants.

B. Scripts

We developed 29 brief scripts for reading: each of which consists of a two sentences for an auditory stimulus and its response. Each script is intended for a specific response concerning emotions, intonation, timing control, or voice volume. The intended responses are categorized into 15 types: focal emphasis, contrastive emphasis, seven emotions (joy, surprise, fear, sadness, anger, avoidance, and disgust), fast/slow speaking rate, immediate/delayed response, and high/low voice volume. Examples of the script are shown in Table. 1. There were two scripts, except for anger (3), avoidance (1), and disgust (1).

C. Auditory Stimuli

To record the auditory stimuli, two actors, one male and one female, read aloud while performing according to the instructions in the text (e.g., Please say "Happy birthday" as you celebrate your close friend's birthday). We finally recorded 29 audio stimuli for each actor (in total, 58 speech samples) with a sampling rate of 44.1 kHz and a quantization precision of 16 bits.

Table 1 Fifteen categories and examples of scripts translated into English (original version in Japanese for each category).

Category	Example of script
Focal emphasis	Where is your hometown? – My hometown is B City.
Contrastive emphasis	Can I get to A City by train? – No, you can't. You should go there by bus.
Joy	Here's a present for you. – Thanks a lot, I've always wanted this.
Surprise	I met Mr. B for the first time in 10 years. – That's surprising, how has he been?
Fear	I heard you had an accident. – I almost died.
Sad	What's happened to your watch? – I've lost it. I liked it, though.
Anger	It's meaningless to do that. – Don't disturb me.
Avoidance	Is this edible? – I don't think so. It stinks.
Disgust	I'm sick of Mr. B. – He is a really unpleasant guy.
Fast speaking rate	What time is the last train? – It's coming, you've only one minute!
Slow speaking rate	It's a really nice loungy place, isn't it? – Yeah, now I am feeling alive again!
Delayed response	This estimate seems too expensive. – It's difficult to get this cheaper anymore.
Immediate response	Your boss is asking and looking for you. – Oh! I'll be there in a minute.
Low volume	This is just between us, but Mr. B is going to leave for another job. – I didn't know that. It's still a secret, isn't it?
High volume	I can't hear you well. – Can you hear me now?

D. Experimental Procedure

The experiment was conducted in a quiet meeting room. A laptop computer with a headset microphone connected and a single-channel loudspeaker ware placed on a table. A participant sat on a chair in front of the table and was instructed to read aloud the script acting according to the situation expected from the presented auditory stimuli. The loudspeaker was placed 50 cm in front of the participant.

In the experiment, 58 response scripts were randomly presented on the display of the laptop PC at the same time as the corresponding auditory stimulus was presented from the loudspeaker. Before the recording of the response, the participant repeated the response script twice for practice. An experimenter pressed a keyboard button to proceed to the next task after the participant finished reading the script. The participant's voice was recorded with the headset microphone synchronizing with the played auditory stimuli to measure the reaction time, at a sampling rate of 44.1 kHz and a quantization precision of 16 bits.

III. Statistical Analyses

A. Frame-Based Speech Feature Extraction

We first extracted fundamental frequency (F_0), intensity, and Mel-cepstral coefficients (MFCCs) frame-by-frame from the recorded sound samples. The MFCCs were obtained and used for the next alignment step. The F_0 and intensity were calculated using Praat [15] and the MFCCs were analyzed using Auditory Toolbox [16] in MATLAB for MFCCs on the basis of the window step of 10 ms. The pitch floor and ceiling for F_0 extraction were set to 75 Hz and 350 Hz for male speakers, and 100 Hz and 500 Hz for the female actor. The logarithm of F_0 values was taken and normalized by subtracting the within-session mean log F_0. The intensity values were normalized using the within-session mean intensity to eliminate the effect of the difference in the recording settings.

B. Alignment using Dynamic Programming (DP) Matching.

The analysis aims are to compare the speech features of responses to the same sentence read aloud by the two actors. Since the pair of responses have different phoneme durations, we should align the time frames of the two utterances of the two actors. We utilized DP [17] matching to align the frames where the distance of their MFCC vectors was close.

Figure 1 shows an example of a path obtained from DP matching that minimizes the cumulative cost. We take the path that minimizes the cost assigned to each path direction. The direction at Point k is selected using the following equation:

$$C(i,j) = \min \begin{bmatrix} C(i-1,j) + d(i-1,j) + d(i,j) \\ C(i-1,j-1) + \frac{4}{3} d(i,j) + d(i-1,j-1) \\ \qquad + d(i-2,j-2) \\ C(i,j-1) + d(i,j-1) + d(i,j) \end{bmatrix}, \quad (1)$$

where $C(i, j)$ and $d(i, j)$ denote the cumulative cost and the distance between two frames at Node (i, j), respectively. The distance $d(i, j)$ is defined by

$$d(i,j) = \frac{1}{w_1 + w_2} \left\{ w_1 \sum_{m=1}^{M-1} \left(x_m^A(i) - x_m^B(j) \right)^2 + w_2 \sum_{m=1}^{M-1} \left(a_m^A(i) - a_m^B(j) \right)^2 \right\}, \quad (2)$$

where $x_m^A(i), x_m^B(i)$ and $a_m^A(i), a_m^B(i)$ are the m-th coefficients of MFCC and ΔMFCC of the i-th frame in Utterances A and B, respectively, and w_1 and w_2 are the weighting factors. After obtaining the best path of DP matching, we compared the prosodic parameters of matched frames of the two speakers. Although the shape of the DP path may also include information about the difference between the responses, we will consider using it as another feature in our future work.

C. Speech Feature Calculations

We measured the following six speech features from each pair of recorded speech samples: one of the pair is the response to a male actor's utterance and the other is that to a female voice. Let a participant's response to the male actor's auditory stimulus on Task k be R_k^m and that to the female actor's one be R_k^f below.

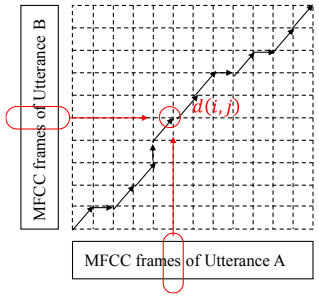

Fig. 1 Search of the best path using DP matching

(1) Utterance-duration ratios

This feature is the time-expansion ratio of the utterance durations of R_k^m to that of R_k^f defined by

$$r_k = \frac{T_k^f}{T_k^m}, \quad (3)$$

where T_k^f and T_k^m denote the utterance durations of R_k^m and R_k^f, respectively.

(2) DP-matching scores

We defined the DP-matching score as the number of analysis frames that matched in the DP path search divided by the total number of frames. If the rhythms of R_k^m and R_k^f are differently produced, the score will be low even when the total durations are the same.

(3·4) Mean absolute errors (MAEs) of log F_0 and intensity

To compare the log F_0 and intensity patterns between two utterances, we measured the MAEs between matched framed between Utterance A and B.

(5·6) 90-percentile values of F_0 and intensity

We adopted the 90-percentile values instead of maximum values, because of their robustness. We measured the difference in 90-percentile values of log F_0 and intensity between R_k^m and R_k^f.

D. Statistical Analyses

For each reading script, we examined whether any prosodic features have a significant difference between the ASD and TD groups by Wilcoxon signed-rank tests ($p < 0.05$). To avoid the increase of Type I errors with multiple comparisons, we corrected p values using Benjamini-Hochberg (BH) adjustment [18] in six features.

E. Classification Analysis

We conducted a classification analysis using the Support Vector Machine (SVM) with the radial kernel implemented in the 'e1071' package in the statistical environment R. A backward feature selection (BFS) method was employed to select an appropriate set of features. In BFS, starting from all 90 features (15 categories × 6 features), the combination with

Table 2 Features that are significantly different between the ASD and TD groups. Medians of each group and adjusted *p*-values of Wilcoxon rank sum tests are also shown.

Category	Feature	ASD	TD	*p*
Contrastive emphasis	90 pctl of F_0	-0.0514	0.0108	0.024
Fear	90 pctl of F_0	-0.0137	0.0430	0.025
Anger	90 pctl of intensity	-0.0879	0.602	0.031
Disgust	MAE of intensity	8.39	5.32	0.029

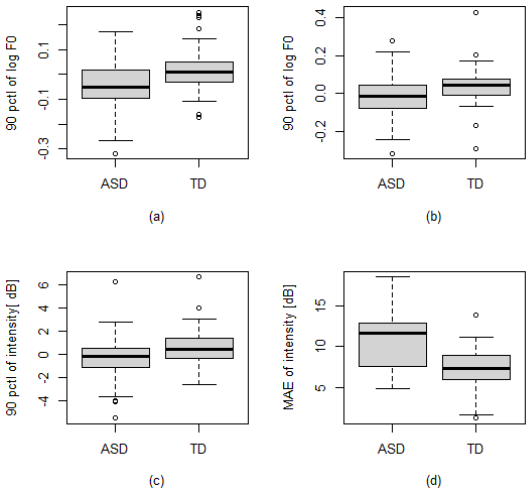

Fig. 2 Distribution of prosodic features of categories 'Contrastive emphasis', 'Fear', 'Anger', and 'Avoidance', in ASD and TD groups

the best classification accuracy was selected by excluding one features at each step. To evaluate the accuracy and *F*-measure, we carried out leave-one-out cross-validation.

IV. Results

A. Comparison between ASD and TD groups

Table 2 shows the prosodic features that had a significant difference between ASD and TD groups. Fig. 2 shows the box plots of prosodic features of each group. In 'Contrastive emphasis' and 'Fear' categories, the 90-percentile value of F_0 was higher for the response to the male actor in the TD group than the ASD group as shown in Fig. 2 ($p < 0.05$). As for 'Anger' category, the 90-percentile value of intensity was higher for the male voice stimuli in the TD group compared to the ASD group ($p < 0.05$). In 'Avoidance' category, the MAE of intensity was higher in the speakers with ASD ($p < 0.05$).

B. Classification Analysis

As a result of the BFS steps, the best accuracy of 90.4% was obtained with five features (the DP-matching scores of categories 'Angry', 'Avoidance', and 'Immediate response', the MAE of the intensity of 'Disgust', and the utterance-duration ratio of 'High voice volume'). The *F*-measure was 0.923 with these features. Note that the prosodic features were

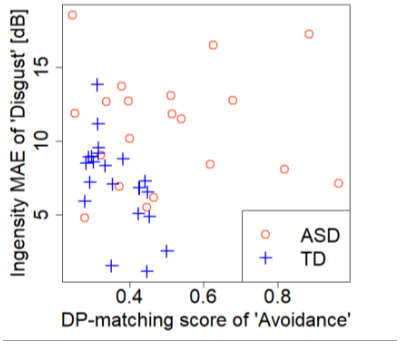

Fig. 3 Scatterplot of two of five features selected in BFS

selected from one from each category. The classification accuracy and *F*-measure were 78.6% and 0.786, respectively, when using the four features with a significant difference between the ASD and TD groups. Fig. 3 shows the scatterplot between the DP-matching scores of 'Avoidance' and the MAE of the intensity of 'Disgust', which had the least *p*-values in the Wilcoxon rank-sum tests in the selected five features. They are separately distributed except that three speakers with ASD and one with TD locate in the distribution of the group other than themselves.

V. Discussions and Conclusions

The participants with TD used a higher maximum pitch in the response to the male actor in the 'Fear' category with a significant difference from those with ASD. Moreover, the median of the ASD group was nearly zero (-0.0137), which indicates that the maximum values were almost the same regardless of the gender of actors. Regarding the 90-percentile values of intensity in the 'Anger' category, the participants with TD also responded with high maximum volume (Difference: 0.6 dB), whereas those with ASD responded with nearly the same volume (Difference: -0.0879 dB). This result is consistent with the study in [13] that revealed the stability of prosodic control in speakers with ASD.

The 90-percentile values of the log F_0 in the 'Contrastive emphasis' category were significantly lower in the ASD group, indicating the tendency of response to the male actor, whereas the TD speakers distributed nearby zero as shown in Fig. 2. This indicates that TD speakers responded similarly to both actors in the contrastive emphasis tasks. This instability of volume control by autistic participants is consistent with the result obtained by Nadig and Shaw who compared the children with high-functioning autism (HFA) and their typical peers [19].

In the 'Disgust' category, the absolute difference between the responses to the two actors was larger in participants with ASD, whereas no significant difference was found in the 90-percentile values of intensity. This indicates that the individuals with ASD did not show the stability that was observed in the two kinds of emotions, 'Fear' and 'Anger'. The study by Zhao *et al.* pointed out the need to investigate the

emotions separately based on the finding of the autistic children's hypervigilance toward 'Disgust' emotion [20].

In the classification analysis with SVM, three of the selected features were related to the rhythm control, and another was related to speaking rate. As shown in Fig. 3, in the 'Avoidance' category, some of the speakers with ASD had high DP-matching scores, indicating that they were more consistently responding to the two actors' voices. This is consistent with the results of stability of duration in children with HFA [19].

In conclusion, the brief scripts developed in this study were effective for quantifying the prosodic characteristics of adults with ASD and successfully provided the set of prosodic features which classify the ASD and TD groups, although further studies of the relationship between the features and severity of ASD should be conducted.

ACKNOWLEDGMENTS

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: 16H01735, 20H00613).

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] S. Y. Bonneh, Y. Levanon, O. Dean-Pardo Lossos, L. Y. Adini, "Abnormal Speech Spectrum and Increased Pitch Variability in Young Autistic Children,". *Frontiers in Human Neuroscience,* vol. 4, pp.237. 2011.

[3] A. Nadig, H. Shaw H, "Acoustic and Perceptual Measurement of Expressive Prosody in High-functioning Autism: Increased Pitch Range and What It Means to Listeners.," *Journal of Autism and Developmental Disorders,* vol. 42, no. 4, pp. 499–511, 2012

[4] M. G. Filipe, S. Frota, S. L. Castro, S. G. Vicente, "Atypical Prosody in Asperger Syndrome: Perceptual and Acoustic Measurements," *Journal of Autism and Developmental Disorders,* vol. 44, no. 8. pp. 1972–1981, 2014.

[5] C. Kaland, E. Krahmer, M. Swerts, "Contrastive Intonation in Autism: The Effect of Speaker-and Listener-perspective," *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[6] Y. Nakai, R. Takashima, T. Takiguchi, S. Takada, "Speech Intonation in Children with Autism Spectrum Disorder," *Brain and Development*, vol. 36, no. 6, pp. 516–522, 2014.

[7] 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.

[8] T. Saga, H. Tanaka, H. Iwasaka, and S. Nakamura, "Objective Prediction of Social Skills Level for Automated Social Skills Training Using Audio and Text Information," In Companion Publication of the 2020 International Conference on Multimodal Interaction, pp. 467-471, 2020.

[9] S. Levy, M. Duda, N. Haber, and D. P. Wall, "Sparsifying Machine Learning Models Identify Stable Subsets of Predictive Features for Behavioral Detection of Autism," *Molecular Autism*, vol. 8, no.1, pp. 1-17, 2017.

[10] M. Kumar, P. Papadopoulos, R. Travadi, D. Bone, and S. Narayanan, "Improving Semi-Supervised Classification for Low-Resource Speech Interaction Applications," *Proceedings of ICASSP 2018*, pp. 5149-5153, pp. 2018.

[11] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. S. Narayanan, "Acoustic-Prosodic Correlates of 'Awkward' Prosody in Story Retellings from Adolescents with Autism," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. S. Narayanan, "Acoustic-Prosodic and Turn-Taking Features in Interactions with Children with Neurodevelopmental Disorders," *Proceedings of Interspeech* 2016, pp. 1185-1189, 2016.

[13] K. Ochi, N. Ono, K. Owada, M. Kojima, M. Kuroda, S. Sagayama, S., 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.

[14] M. Kissine and P. Geelhand, "Brief report: Acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder," *Journal of autism and developmental disorders,* vol. 49, no. 6, pp. 2572-2580, 2019.

[15] P. Boersma, "Praat, a System for Doing Phonetics by Computer," *Glot International* vol. 5 no. 9/10, pp. 341-345, 2001.

[16] S. Malcolm, "Auditory Toolbox Version 2," https://engineering. purdue. edu/~ malcolm/interval/1998-010/

[17] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Trans., ASSP*, vol. 26, no. 1, pp. 43-49, 1978.

[18] 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.

[19] A. Nadig, and H. Shaw, "Acoustic Marking of Prominence: How Do Preadolescent Speakers with and without High-Functioning Autism Mark Contrast in an Interactive Task?," *Language, Cognition and Neuroscience,* vol. 30, no. 1-2, pp. 32-47, 2015.

[20] X. Zhao, P. Zhang, L. Fu, and J. H. Maes, "Attentional Biases to Faces Expressing Disgust in Children with Autism Spectrum Disorders: an Exploratory Study," *Scientific Reports*, vol. 6, no. 1, pp. 1-9, 2016.