Vocal analysis of speech in adults with autism spectrum disorders

I-Fan Lin¹,²,³*, Sadao Hiroya¹, Kosuke Asada⁴, Satsuki Ayaya⁴, Shin-ichiro Kumagaya⁴ and Masaharu Kato¹,⁵

¹Human Information Science Laboratory, NTT Communication Science Laboratories, 3–1, Morinosato Wakamiya, Atsugi, 243–0198 Japan
²National Taiwan University Hospital, No. 7, Zhongshan South Road, Taipei, 100, Taiwan
³National Health Research Institutes, No. 35, Keyan Road, Zhunan, Miaoli, 350, Taiwan
⁴Research Center for Advanced Science and Technology, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo, 153–0041 Japan
⁵Center for Baby Science, Doshisha University, 4–1–1 Kizugawadai, Kizugawa, 619–0225 Japan

(Received 1 June 2017, Accepted for publication 5 October 2017)

Keywords: Autism, Prosody, Vocal analysis, Speech rhythm

PACS number: 43.70.Dn  [doi:10.1250/ast.39.154]

1. Introduction

Autism spectrum disorders (ASD) has been defined as a group of conditions in which the core impairments are in two main areas: (1) social interaction and communication and (2) restricted and repetitive behaviors. In oral communication, prosody, including intonation, tone, stress, and rhythm, is important for us to understand the emotional state of the speaker and to differentiate the intention of the utterances. However, prosodic errors have been noticed in individuals with ASD since the early days of ASD research [1]. Recent studies reveal abnormal loudness, pitch, and stress in ASD individuals [2,3]. Automatized acoustic analysis of spontaneous vocal utterances in ASD children also shows reduced developmental projections of speech production such as pitch control and speech durations [4].

This study measured speech fluency, short pause duration, and durations of individual phonemes. We hypothesized that ASD adults who could speak fluently might still have residual errors in speech rhythm since they have been found to have residual phonetic and phonological errors and prosodic errors. In addition, we also analyzed formant frequencies in their vowel pronunciations.

2. Material and methods

2.1. Participants

Ten participants with ASD and thirteen neurotypical (NT) participants joined this study, and they were matched by age and intelligence quotient (IQ). Table 1 provides detailed information about the participants. Verbal IQ (VIQ) and performance IQ (PIQ) were measured by WAIS-III or WAIS-R [6]. The participants with ASD were recruited through a self-help group for people with developmental disabilities, and their ASD diagnosis was provided by hospitals based on DSM-IV-TR [7]. All except one ASD participants received Autism Diagnostic Observation Schedule (ADOS) [8] evaluation. Five in nine ASD participants fit the criteria for autism spectrum based on ADOS (module 4) evaluation (communication score ≥ 2 and social interaction ≥ 4, and communication + social interaction ≥ 7).

Written consent was obtained from all participants before we conducted the experiments. All procedures were conducted in accordance with the Declaration of Helsinki and approved by the ethics committee of NTT Communication Science Laboratories. All of the participants were naive to the purposes of the study. The participants were paid for their time.

2.2. Apparatus

All auditory experiments were conducted in a sound-insulated booth and controlled by a computer. The participants’ speech was collected with a microphone (Rode NT2-A). A mixer (Soundcraft EPM6) was used to send the collected speech to a noise filter (Behringer Multigate Pro XR4400) to filter out background noise. The filtered speech signals were sent to an audio interface (Roland EDIROL UA-101) and a computer for recording. The sampling rate for recording was 48,000 Hz.

2.3. Procedures

We selected the Japanese version of the short article listed in a previous paper for testing speech information rate [9]. The participants practiced three times before the recording. They were asked to speak as fluently as possible, and they were not instructed to speak with a certain accent. There was no feedback for their reading.

We decided the start and end of a phoneme by both listening to the speech and looking at its speech waveform and spectrogram. Regarding voice onset time (VOT) for voiced and unvoiced consonants, the start and end of it was decided by looking at the spectrogram carefully to find the gap between release and voicing. For formant analysis for vowels, we used a robust method against fundamental frequency [10].

*e-mail: ifanlintw@gmail.com
There were 22 phonemes presented in this reading material. Averaged phoneme duration for each phoneme was computed for each participant. There was no significant between-group duration difference in most phonemes, including vowels, except that the duration of /t/ and /k/ was significantly longer in the ASD group than in the NT group ($t(15.24) = 4.25$, $p < 0.001$, $r = 0.68$ and $t(18.92) = 4.09$, $p < 0.001$, $r = 0.67$, respectively) and the duration of /d/, /g/, and /r,l/ was significantly shorter in the ASD group than in the NT group ($t(14.54) = 3.89$, $p < 0.001$, $r = 0.65$ and $t(20.03) = 2.28$, $p = 0.03$, $r = 0.45$ and $t(16.01) = 3.49$, $p = 0.003$, $r = 0.61$, respectively) (For the details, please see Table 2). Regarding these five phonemes, /t/, /d/, /k/, /g/, /r,l/ were presented 26/16/2, 10/8/2, 16/7/2, 5/4/1, and 13/1/2 times (presented in the format of ‘total number’/‘number in the initial syllable position’/‘number in the accented mora’) in this reading material. On the other hand, /p/ and /b/ were not included in this analysis because /b/ was only presented once, and /p/ was absent in this reading material.

3.4. Voice onset time (VOT)

During phoneme labelling for duration computation, the beginning of the consonants /t,d,k,g,ch/ was set to be the end of their previous phoneme. To further investigate motor control for the consonants /t,d,k,g/, we analyzed their VOTs. The averaged VOTs of /t/ and /d/ in the ASD group were 16 and 14 ms, and the averaged VOTs of /t/ and /d/ in the control group were 19 and 14 ms. The averaged VOTs of /k/ and /g/ in the ASD group were 37 and 23 ms, and the averaged VOTs of /k/ and /g/ in the control group were 35 and 26 ms. There was no significant between-group difference in the VOT of /t/ ($t(17.96) = 1.27$, $p = 0.22$, $r = 0.27$), /d/ ($t(19.64) = 0.29$, $p = 0.78$, $r = 0.06$), /k/ ($t(17.10) = 0.91$, $p = 0.37$, $r = 0.20$), or /g/ ($t(20.97) = 0.90$, $p = 0.38$, $r = 0.19$).

3.5. Phoneme errors

There was only one simple pronunciation error observed in one participant. No participant spoke additional phonemes. On average, ASD participants missed 3.54 phonemes and NT participants missed 3.54 phonemes. There was no significant between-group difference in the number of missed phonemes ($t(18.54) = 1.22$, $p = 0.24$, $r = 0.26$) or in the total phoneme errors ($t(20.73) = 0.45$, $p = 0.10$, $r = 0.24$).

3.6. Fundamental frequency and formants

Since the f0 and formants usually differ between genders, we analyzed the f0, first formant ($F1$), and second formant ($F2$) in only female participants (eight in each group). The mean and standard deviation of $f0$, $F1$, and $F2$ across the voiced intervals for each vowel were calculated (Table 3). Six two-way mixed model ANOVAs were conducted for the mean and standard deviation of $f0$, $F1$, and $F2$ with Group (ASD and NT) as the between-subject factor and Vowel (/a/, /e/, /i/, /o/ , /u/) as the within-subject factor.

For the mean of f0, $F1$, and $F2$, Vowel showed a significant effect ($F(4, 56) = 28.06$, 64.09, 601.59, respectively and, $p < 0.001$ for all of them, partial $\eta^2 = 0.67$, 0.82 and 0.98, respectively), but neither Group ($F(1, 144) = 0.02$, $p < 0.001$, $r = 0.67$, respectively) and the duration of /d/, /g/, and /r,l/ was significantly shorter in the ASD group than in the NT group ($t(14.54) = 3.89$, $p < 0.001$, $r = 0.65$ and $t(20.03) = 2.28$, $p = 0.03$, $r = 0.45$ and $t(16.01) = 3.49$, $p = 0.003$, $r = 0.61$, respectively) (For the details, please see Table 2). Regarding these five phonemes, /t/, /d/, /k/, /g/, /r,l/ were presented 26/16/2, 10/8/2, 16/7/2, 5/4/1, and 13/1/2 times (presented in the format of ‘total number’/‘number in the initial syllable position’/‘number in the accented mora’) in this reading material. On the other hand, /p/ and /b/ were not included in this analysis because /b/ was only presented once, and /p/ was absent in this reading material.
The fundamental frequencies ($f_0$) and first and second formant frequencies ($F_1$ and $F_2$) (in Hz) in the ASD and NT groups, expressed as mean (standard deviation).

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD $f_0$</td>
<td>201(32)</td>
<td>199(32)</td>
<td>202(32)</td>
<td>198(30)</td>
<td>211(26)</td>
</tr>
<tr>
<td>ASD $F_1$</td>
<td>643(105)</td>
<td>444(45)</td>
<td>365(48)</td>
<td>462(71)</td>
<td>371(36)</td>
</tr>
<tr>
<td>ASD $F_2$</td>
<td>616(94)</td>
<td>484(53)</td>
<td>397(36)</td>
<td>470(76)</td>
<td>404(30)</td>
</tr>
<tr>
<td>NT $f_0$</td>
<td>202(20)</td>
<td>201(20)</td>
<td>204(20)</td>
<td>198(22)</td>
<td>215(20)</td>
</tr>
<tr>
<td>NT $F_1$</td>
<td>639(171)</td>
<td>439(92)</td>
<td>359(59)</td>
<td>463(73)</td>
<td>373(36)</td>
</tr>
<tr>
<td>NT $F_2$</td>
<td>1725(134)</td>
<td>2279(149)</td>
<td>2394(155)</td>
<td>1367(63)</td>
<td>1813(119)</td>
</tr>
</tbody>
</table>

0.58 and 0.00, $p = 0.90$, 0.46 and 0.96, partial $\eta^2 = 0.00$, 0.04 and 0.00, respectively) nor their interaction did ($F(4, 56) = 0.35, 1.24$ and 0.38, $p = 0.75, 0.30$ and 0.80, partial $\eta^2 = 0.02, 0.08$ and 0.03, respectively).

For the standard deviation of $f_0$, $F_1$, and $F_2$, Vowel showed a significant effect ($F(4, 56) = 5.79, 6.01$ and 9.65 and $p = 0.01, 0.01$ and $<0.001$, partial $\eta^2 = 0.29, 0.30$ and 0.41, respectively), but neither Group ($F(1, 14) = 0.49, 0.88$ and 0.31 and $p = 0.50, 0.36$ and 0.59, partial $\eta^2 = 0.03, 0.06$ and 0.02, respectively) nor their interaction did ($F(4, 56) = 1.38, 0.54$ and 1.28 and $p = 0.27, 0.58$ and 0.29, partial $\eta^2 = 0.09, 0.04$ and 0.08, respectively).

In summary, there was no significant between-group difference for the mean and standard deviation of the $f_0$ and the frequencies of the first two formants.

4. Discussion

This paper examined speech in well-controlled speech materials of adults with and without ASD. The results show that ASD participants had longer phoneme duration for unvoiced consonants /t/ and /k/ and shorter phoneme duration for voiced consonants /d/, /g/, and /r/. Otherwise, there was no significant between-group difference in speech fluency, pause durations, voice onset times (for /d/, /g/, /t/, and /k/), phoneme errors, fundamental frequency or formants.

The finding of no significant between-group difference in speech fluency observed in this study is consistent with previous studies [2,12,13], which support the argument that ASD individuals do not have speech motor disorders. On the other hand, the significant between-group difference in the duration of these four consonants, /d/, /g/, /t/, and /k/, indicates that ASD adults might still have some residual errors due to audio-vocal control. Feedback control is argued to play an important role for speech timing control [14,15], and ASD individuals have been found to over rely on auditory feedback in pitch and timing control [16,17].

This study has some limitations. Syllable position, accent condition, and dialects spoken by the speakers all have an effect on phoneme duration and VOT. The effect of syllable position and accent condition was not analyzed in this study because there was not sufficient sample size in some categories for the phonemes mentioned above. The effect of dialect was not analyzed, either, because we did not have this information.

Future studies should include a larger group of participants to evaluate the effect of gender, age, and symptom severity. More specifically, the male/female ratio in autism is usually found to be larger than 1, but we had more female than male participants due to our recruitment methods. This gender difference further led us to emphasize on the formant frequency analysis in our female participants rather than our male participants because the sample size of the male group was not big enough for robust statistical analysis. It would also be valuable if future studies could evaluate speech rhythm in long and natural utterances for more detailed analysis. Moreover, it would be informative if future studies could find the correlation between the vocal analysis results of their speech and the perceptual impressions of it.

References


