PERCEPTION OF JAPANESE PITCH ACCENT BY TYPICALLY DEVELOPING CHILDREN AND CHILDREN WITH AUTISM SPECTRUM DISORDERS

Yasuaki Shinohara[1-3], Mariko Uchida[4,5], Tomoko Matsui[4]

[1]Waseda University, [2]CUNY Graduate Center, [3]University of Delaware, [4]Chuo University, [5] Tokyo Gakugei University
y.shinohara@waseda.jp

ABSTRACT

This study examined how typically developing (TD) children and children with autism spectrum disorders (ASD) identified Japanese pitch-accent minimal-pair words in three stimulus conditions: natural recordings (NatRec), sine-wave speech (SWS), and noise-vocoded sine-wave speech (NzVocSWS). It was hypothesized that TD children would show a significant decrease in identification accuracy from the NatRec condition to degraded speech (SWS and NzVocSWS) conditions, and a significant increase from the SWS to NzVocSWS condition, consistent with the results in a previous study testing Japanese-speaking adults. On the contrary, children with ASD were expected to show a different identification pattern because of their difficulty in phonological processing and detecting subtle differences in duration and intensity cues. The results demonstrated that both groups showed a similar pattern, although the identification accuracy of children with ASD was significantly lower than that of TD children across the three stimulus conditions.

Keywords: speech perception, pitch accent, autism spectrum disorders, noise-vocoded sine-wave speech

1. INTRODUCTION

Speech sounds carry multiple acoustic cues signaling phonological contrasts [1–4], and degraded speech is often used to manipulate audible acoustic cues. Sine-wave speech (SWS) is a type of synthetic speech, consisting of multiple time-varying sinusoids that replicate the first three or four formants of natural speech with its amplitude pattern [5]. People understand words and sentences from SWS [5–7], but since there is no fundamental frequency (F0) information in SWS, Japanese speakers cannot identify pitch-accent minimal-pair words correctly [8]. Instead of F0, they rely on the first formant frequency (F1) to identify pitch accents, resulting in poor identification [9]. However, when SWS is noise-vocoded, individual formants become less audible, and Japanese speakers use other acoustic cues (e.g., duration and intensity), leading to better identification [9]. In the present study, pitch-accent perception of children with autism spectrum disorders (ASD) was examined with the hypothesis that their pitch-accent perception would be different than typically developing (TD) children.

ASD is often described as a neurodevelopmental disorder affecting social interaction and speech communication [10]. People with ASD often have enhanced pitch perception, but their outperformance is observed on a low-level perception (e.g., auditory discrimination), and not on a higher-level (e.g., categorical) perception [11–13]. This means that people with ASD are auditorily sensitive, but experience difficulty in phonological processing. This claim has been supported by both behavioral and neurophysiological experiments [12].

The Japanese language has lexical contrasts based on a high (H) vs. low (L) bitonal pitch accent. For example, [niꜜzi] (HL) with a relatively high-pitched first mora [ni] and a relatively low-pitched second mora [zi] means *two o'clock*, whereas [niꜛzi] (LH) with a low-pitched first mora and a high-pitched second mora means *rainbow*. The primary acoustic cue for pitch-accent contrast is F0, and Japanese speakers use the location of the F0 peak to identify accented morae [1]. Although there are many debates about whether duration and intensity are the secondary acoustic cues for pitch-accent perception [2–4], a previous study showed that Japanese speakers tend to rely on the duration and intensity of vowels to identify pitch-accent minimal-pair words when F0 is not available [9].

This study examined the identification accuracy of Japanese pitch-accent minimal-pair words in three testing conditions: natural recordings (NatRec), SWS, and noise-vocoded sine-wave speech (NzVocSWS). In NatRec, where all relevant acoustic cues, such as F0, duration, and intensity, are available, it was hypothesized that the identification accuracy would be at ceiling for TD children. However, in SWS, where F0 is not available, TD children may rely on F1 for pitch-accent perception, as Japanese-speaking adults do [9]. Consequently, TD children would not

be able to identify pitch accents correctly. In NzVocSWS, where the effect of quasi-periodicity (i.e., F1) is reduced, TD children were expected to increase their identification accuracy compared to the SWS condition. Since each formant is less audible due to the noise-vocoding effect, they may rely on other acoustic cues (e.g., duration and intensity), resulting in higher identification accuracy [9]. Conversely, it was hypothesized that children with ASD would not increase their identification accuracy from SWS to NzVocSWS due to their difficulty in phonological processing and detecting subtle differences in both duration and intensity cues [12], [14–19]. For children with ASD, identification accuracy was expected to be at a chance level in both the SWS and NzVocSWS conditions, and not as high as that of TD children in the NatRec condition. Based on these hypotheses, we predicted a significant interaction between the group and condition factors.

2. METHOD

2.1. Participants

Table 1 displays the participants' information. Sixteen children participated in the experiment. They were all Tokyo Japanese speakers who had lived most of their lives in Tokyo [20]. Eight of the 16 participants were TD children, and the remaining eight were diagnosed by pediatricians as having ASD. The two groups (TD and ASD) were matched in chronological age, but significantly differed in the scores of the Parent-Interview ASD Rating Scales – Text Revision (PARS-TR) [21]. The PARS-TR is an interview-based rating scale widely used in Japan to assess ASD symptoms. The symptoms were assessed during the preschool period and at the time of the study.

	TD (n = 8) Mean (SD)	ASD (n = 8) Mean (SD)	Group differences
Sex	6 girls, 2 boys	2 girls, 6 boys	
Age (years)	9.64 (1.74)	9.48 (1.74)	n.s. ($p > .05$)
PARS-TR (present)	4.00 (6.76)	20.88 (9.72)	$p < .01$
PARS-TR (preschool)	7.00 (10.62)	30.88 (10.87)	$p < .01$
PVT-R score	63.38 (15.58)	46.50 (23.81)	n.s. ($p > .05$)
PVT-R (SS)	13.63 (3.93)	8.88 (6.22)	$p = .089$

Note: Independent sample t-test and Wilcoxon rank sum test were used for the group difference analyses.

Table 1: Participants' information of typically developing (TD) children and children with autism spectrum disorders (ASD).

There was no significant difference in the Peabody Picture Vocabulary Test of receptive vocabulary (Japanese version; PVT-R score) between the two groups, but the difference in the PVT-R evaluation score (SS) was marginally significant. The PVT-R score is often used to assess verbal intelligence quotient (IQ) [22], and the PVT-R (SS) is often used to assess language delay [23].

2.2. Stimuli

Three types of stimuli were used in our experiment: NatRec, SWS, and NzVocSWS. Figure 1 displays waveforms and narrowband spectrograms of the word [ame] with the HL and LH pitch accents in each stimulus type.

2.1.1. Natural recordings (NatRec)

Two Japanese pitch-accent minimal pairs ([ame] with HL meaning *rain* vs. LH meaning *candy* and [niꜜzi] with HL *two o'clock* vs. LH *rainbow*) were recorded by six Tokyo Japanese speakers (three women and three men), using a Rode NT2-A microphone with 44,100 16-bit samples per second. The intensity was normalized across all tokens based on the root-mean-square method. Tokens from two speakers (two women) were used only for practice sessions, and the tokens from the remaining four speakers were used for the identification test. Sixteen tokens (4 minimal-pair words × 4 speakers) were used in the identification test of the NatRec condition.

2.1.2. Sine-wave speech (SWS)

The duration of the NatRec stimuli was averaged between minimal-pair words produced by each speaker. All tokens were then transformed to SWS with a Praat script written by Darwin [24]. The first three formants (F1, F2, and F3) and their amplitude were tracked every 10 ms with the respective gender setting. Other parameters were set as the default (e.g., amplitude tracks were low-pass filtered at 50 Hz). The three individual sinusoids were combined to construct the SWS version of the 16 tokens (4 minimal-pair words × 4 speakers).

2.1.3. Noise-vocoded sine-wave speech (NzVocSWS)

Finally, the SWS stimuli went through a peak-picking noise vocoder using a Praat script written by Winn [25]. The frequency range was set from 70 Hz to 10 kHz. Spectral analysis and synthesis were conducted using 33 filters, with the width of the synthesis filter being 6 dB per mm. As each formant was less audible, it was expected that TD Japanese speakers would not rely on F1 but would instead use duration and

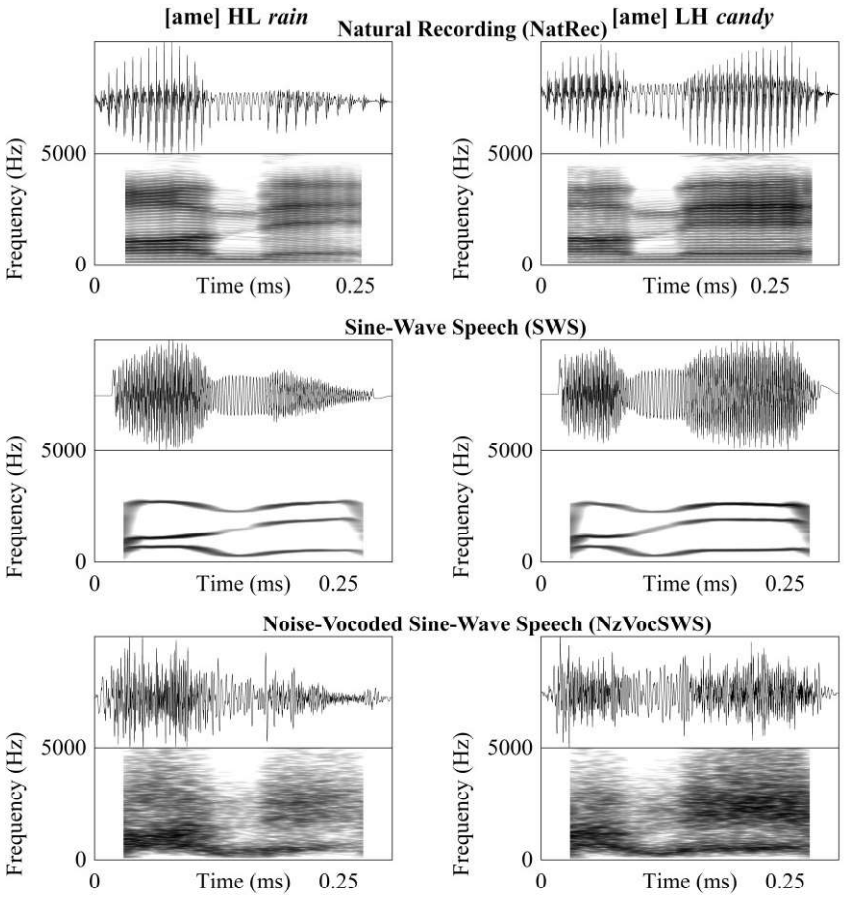

Figure 1: Waveforms and narrowband spectrograms of the word [ame] with the HL (high-low) and LH (low-high) pitch accents in three stimulus types: natural recordings (NatRec), sine-wave speech (SWS), and noise-vocoded sine-wave speech (NzVocSWS).

intensity for pitch-accent perception in the NzVocSWS condition.

It should be noted that when a noise vocoder with square-shaped synthesis filters was used, Japanese speakers did not show a significant difference in their pitch-accent identification accuracy between the SWS and NzVocSWS conditions [8].

2.3. Procedure

Participants took a Japanese pitch-accent identification test under three conditions (NatRec, SWS, NzVocSWS). They heard a word and saw a minimal pair displayed on the screen with both *kanji* (a Japanese writing system with Chinese characters) and *hiragana* (Japanese moraic orthography) (e.g., 雨(あめ), 飴(あめ)), then clicked on the word they thought they had heard. Half of the participants took the test in the NatRec, SWS, and NzVocSWS order, and the other half took it in the NatRec, NzVocSWS, and SWS order. Before each condition, participants had a practice session with two tokens. They received no feedback during the test and heard eight tokens for each minimal pair of [ame] and [niʑi]. Each participant completed 48 trials (8 tokens × 2 minimal pairs × 3 conditions).

3. RESULTS

Figure 2 displays the identification accuracy of pitch-accent contrast under three stimulus conditions (NatRec, SWS, NzVocSWS) by TD children and children with ASD. A logistic mixed effects model based on correct and incorrect binomial responses was used for the statistical analysis. The fixed factors were group (TD, ASD), condition (NatRec, SWS, NzVocSWS), and their interaction. Orthogonal contrasts were set for these categorical variables. By-participant random intercepts were also included in the model.

The logistic mixed effects model demonstrated a significant effect of group (TD vs. ASD), $\beta = 0.32$, $SE = 0.15$, $z = 2.11$, $p = .035$, suggesting that the TD children had higher identification accuracy across all three conditions than the children with ASD. Significant effects were also found for condition. The identification accuracy was significantly higher in NatRec than in degraded speech (i.e., SWS and NzVocSWS) conditions, $\beta = 0.70$, $SE = 0.08$, $z = 8.82$, $p < .001$, and accuracy in the NzVocSWS condition was significantly higher than that in SWS, $\beta = -0.28$, $SE = 0.09$, $z = -3.05$, $p < .01$. However, the interaction

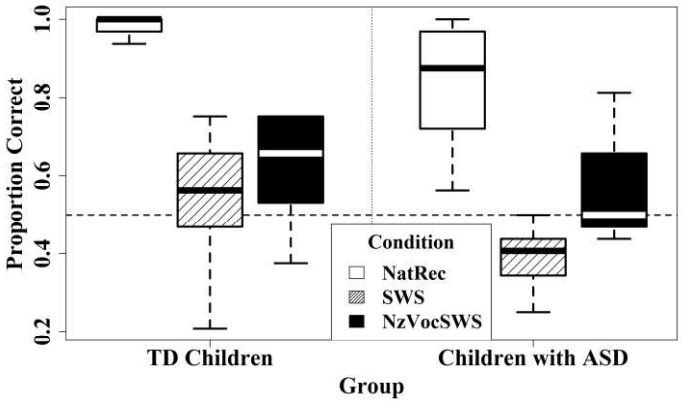

Figure 2: Identification accuracy of two Japanese pitch-accent minimal pairs ([ame] HL *rain* vs. LH *candy*, [niꜜzi] HL *two o'clock* vs. LH *rainbow*) under three stimulus conditions (NatRec, SWS, NzVocSWS) by typically developing (TD) children and children with autism spectrum disorders (ASD). The horizontal dashed line represents the chance level of correct proportion.

between group and condition was not significant in any contrast ($p > .05$).

4. DISCUSSION

This study examined identification accuracy of Japanese pitch-accent minimal-pair words by TD children and children with ASD under three stimulus conditions (NatRec, SWS, NzVocSWS). For TD children, it was hypothesized that pitch-accent identification accuracy would decrease from the NatRec to degraded speech (SWS and NzVocSWS) conditions, but significantly increase from the SWS to NzVocSWS condition. As Japanese-speaking adults did in a previous study [9], TD children may rely on F1 for pitch accent perception in SWS where F0 information is not available, resulting in poor identification accuracy. However, because the effect of misleading voice pitch contour (i.e., F1) may be reduced due to less audible formants in NzVocSWS, we predicted that their pitch accent identification accuracy would increase, with more reliance on other acoustic cues (e.g., duration and intensity). In contrast to TD children, children with ASD were not expected to significantly increase their identification accuracy even after the SWS was noise-vocoded, due to their difficulty in phonological processing and detecting subtle differences in duration and intensity cues [14–19]. Because of these difficulties, identification accuracy in the NatRec condition may also not be as high for children with ASD, predicting a significant interaction between the group and condition factors.

The results demonstrated that, although identification accuracy was lower for the ASD group than the TD group across all three conditions, no significant interaction was found—both groups showed similar pitch-accent perception patterns. Their identification accuracy in the degraded speech (SWS and NzVocSWS) conditions was lower than that in the NatRec condition, and accuracy in the NzVocSWS condition was higher than in the SWS condition. These results do not fully support our hypotheses, but are consistent with the results of tests with Japanese-speaking adults. Since this study did not examine how each acoustic cue is used by each group, further research is necessary.

There are multiple limitations that need to be addressed in a future study. First, language ability was not controlled between the TD and ASD groups. There were no significant differences in the PVT-R (SS) or PVT-R score between the TD and ASD groups, which suggests that the ASD group had neither significant language delay nor significantly lower verbal IQ compared to the TD group. Since autism spectrum disorder is an umbrella term including a wide range of disorders, more-detailed language-ability backgrounds should be analyzed. Second, since these are preliminary results from only eight participants in each group, there may have been a lack of statistical power to show a significant interaction between group and condition—it is necessary to increase the number of participants. Finally, this study only examined pitch-accent identification, not auditory discrimination. It is also important to test auditory sensitivity differences between the two groups. Identifying difficulties in pitch-accent perception in further studies would help clinical diagnoses and education practices for these populations.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers 19H01753 and 23K00490. We thank our participants and research assistants (Ms. Juri Fukuda, Mr. Takumi Fujimura, and Mr. Satsuki Kurokawa) for their support.

6. REFERENCES

[1] Kitahara, M. 2001. *Category Structure and Function of Pitch Accent in Tokyo Japanese*. Indiana University.

[2] Sugiyama, Y. 2022. Identification of minimal pairs of Japanese pitch accent in noise-vocoded speech. *Front. Psychol.* 31. doi: 10.3389/fpsyg.2022.887761

[3] Beckman, M. E., Pierrehumbert, J. B. 1986. Intonational structure in Japanese and English. *Phonol. Yearb.* 3, 255–309. doi: 10.1017/S095267570000066X.

[4] Cutler, A., Otake, T. 1999. Pitch accent in spoken-word recognition in Japanese. *J. Acoust. Soc. Am.* 105, 1877–1888. doi: 10.1121/1.426724

[5] Remez, R. E., Rubin, P. E., Pisoni, D. B., Carrell, T. D. 1981. Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191.

[6] Hillenbrand, J. M., Clark, M. J., Baer, C. A. 2011. Perception of sinewave vowels. *J. Acoust. Soc. Am.* 129, 3991–4000. doi: 10.1121/1.3573980

[7] Remez, R. E., Dubowski, K., Broder Hytowitz, R., Davids, M., Grossman, Y. S., Moskalenko, M., Pardo, J. S., Hasbun, S. M. 2011. Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 968–977. doi: 10.1037/a0020734.

[8] Shinohara, Y. 2022. Perception of noise-vocoded sine-wave speech of Japanese pitch-accent words. *JASA Express Lett.* 2, 085204. doi: 10.1121/10.0013423.

[9] Shinohara, Y. 2022. Japanese pitch-accent perception of noise-vocoded sine-wave speech. *J. Acoust. Soc. Am.* 152, A175. doi: 10.1121/10.0015940.

[10] Haesen, B., Boets, B., Wagemans, J. 2011. A review of behavioural and electrophysiological studies on auditory processing and speech perception in autism spectrum disorders. *Res. Autism Spectr. Disord.* 5, 701–714. doi: 10.1016/j.rasd.2010.11.006.

[11] Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., Bonnel, A.-M. 2003. Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *J. Cogn. Neurosci.* 15, 226–235. doi: 10.1162/089892903321208169.

[12] Wang, X., Wang, S., Fan, Y., Huang, D., Zhang, Y. 2017. Speech-specific categorical perception deficit in autism: An event-related potential study of lexical tone processing in Mandarin-speaking children. *Sci. Rep.* 7, 43254. doi: 10.1038/srep43254.

[13] Hudry, K. et al. 2010. Preschoolers with autism show greater impairment in receptive compared with expressive language abilities. *Int. J. Lang. Commun. Disord.* 45, 681–690. doi: 10.3109/13682820903461493.

[14] Bruneau, N., Bonnet-Brilhault, F., Gomot, M., Adrien, J.-L., Barthélémy, C. 2003. Cortical auditory processing and communication in children with autism: electrophysiological/behavioral relations. *Int. J. Psychophysiol.* 51, 17–25. doi: 10.1016/S0167-8760(03)00149-1.

[15] Bruneau, N., Roux, S., Adrien, J. L., Barthélémy, C. 1999. Auditory associative cortex dysfunction in children with autism: evidence from late auditory evoked potentials (N1 wave–T complex). *Clin. Neurophysiol.* 110, 1927–1934. doi: 10.1016/S1388-2457(99)00149-2.

[16] Isaksson, S., Salomäki, S., Tuominen, J., Arstila, V., Falter-Wagner, C. M., Noreika, V. 2018. Is there a generalized timing impairment in Autism Spectrum Disorders across time scales and paradigms? *J. Psychiatr. Res*., 99, 111–121. doi: 10.1016/j.jpsychires.2018.01.017.

[17] Kargas, N., López, B., Reddy, V., Morris, P. 2015. The relationship between auditory processing and restricted, repetitive behaviors in adults with autism spectrum disorders, *J. Autism Dev. Disord*., 45, 658–668. doi: 10.1007/s10803-014-2219-2.

[18] Vlaskamp, C. et al. 2017. Auditory processing in autism spectrum disorder: Mismatch negativity deficits, *Autism Res*., 10, 1857–1865, doi: 10.1002/aur.1821.

[19] Lepistö, T., Kujala, T., Vanhala, R., Alku, P., Huotilainen, M., Näätänen, R. 2005. The discrimination of and orienting to speech and non-speech sounds in children with autism. *Brain Res*. 1066, 147–157. doi: 10.1016/j.brainres.2005.10.052.

[20] Akinaga, K., Kindaichi, H. 2014. *Shin Meikai Nihongo Akusento Jiten*, 2nd edition. Sanseido.

[21] Pervasive Developmental Disorders Autism Society Japan Rating Scale (PARS) Committee. 2013. *Pervasive Developmental Disorders Autism Society Japan Rating Scale: Text Revision*. Spectrum Publishing Co.

[22] Ueno, K., Nagoshi, N., Konuki, S. 2008. *PVT-R kaiga goi hattatsu kensa. [Picture Vocabulary Test]*. Nihon Bunka Kagakusha.

[23] Kjelgaard, M. M., Tager-Flusberg, H. 2001. An investigation of language impairment in autism: Implications for genetic subgroups. *Lang. Cogn. Process.* 16, 287–308. doi: 10.1080/01690960042000058.

[24] Darwin, C. 2005. SWS [Praat Script]. http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS.

[25] Winn, M. 2021. Vocoder [Praat Script]. http://www.mattwinn.com/praat/vocode_all_selected_v45.txt.