NEUTRALIZATION AND SECONDARY ACOUSTIC CUES OF VOICING CONTRAST: A TOHOKU AND TOKYO JAPANESE PRODUCTION EXPERIMENT

Chuyu Huang[1], Sanae Matsui[2], Naoya Watabe[3], Hiroto Noguchi[2, 4], Ayako Hashimoto[5], Ai Mizoguchi[6, 7], Mafuyu Kitahara[2]

[1]Nagoya Gakuin University [2]Sophia University [3]The University of Tokyo [4]Tokyo Medical and Dental University [5]Tokyo Kasei Gakuin University [6]Maebashi Institute of Technology [7]National Institute for Japanese Language and Linguistics

huang@ngu.ac.jp, sanaematsui1107@gmail.com, watabe.naoya.2821@goo.jp, noguchih425@gmail.com, hassy@kasei-gakuin.ac.jp, aimizoguchi@maebashi-it.ac.jp, mafuyu@sophia.ac.jp

ABSTRACT

This study investigates the phonological voicing contrast of plosives in two Japanese dialects: Tohoku Japanese, which is known for its intervocalic voicing of voiceless plosives, and Tokyo Japanese. Despite some previous research on this topic, the extent to which various acoustic properties, such as voice onset time (VOT) and fundamental frequency (F0) (1) covary with intervocalic voicing in Tohoku Japanese and (2) compare to Tokyo Japanese remains unclear. Thus, we conducted online production experiments to examine the voicing contrast in Tohoku Japanese. Our results indicate that the voicing contrast in Tohoku Japanese is incomplete; moreover, other acoustic cues such as F0 and preceding vowel duration, vary between Tohoku and Tokyo Japanese. These findings suggest that while secondary cues, i.e. duration and onset F0 may play an active role in realizing the intervocalic voicing contrast, cue utilization may differ between these two dialects.

Keywords: VOT, neutralization, Tohoku dialects, secondary cues, onset F0.

1. INTRODUCTION

Tohoku Japanese is a group of dialects of Japanese spoken mainly in the North-Eastern regions of Japan, in which voiceless plosives /t, k/ have been typically realized as voiced [d, ɡ] in intervocalic environments [1–4]. The phonologically voiced plosives /d, ɡ/ in some of these dialects were realized as nasalized plosives; however, in the variations where /d, ɡ/ remain oral plosives, voiceless and voiced plosives may be merged due to the intervocalic voicing.

In addition to the internal change of the Tohoku dialect, Tokyo Japanese, the dominant dialect in Japan, also influences the realization of Tohoku plosives. Both dialects may share common undergoing changes. Tokyo Japanese has undergone a significant change wherein initial voiced consonants are realized with non-negative voice onset times (VOTs) [5]. On the other hand, Tohoku Japanese also exhibits a gradual change, with positive VOTs observed in initial voiced segments. Moreover, the younger generation in Tohoku shows a stronger tendency towards positive VOTs in initial voiced segments [6].

VOT, however, does not seem to be the only acoustic cue used for distinguishing the voicing contrast. It has been noted that Tokyo Japanese speakers utilize other phonetic cues in voicing contrast, such as F0 of the following vowel, to realize or detect the word-initial voicing contrast despite the partial neutralization of the primary VOT cue [7, 8]. Meanwhile, research of Tohoku dialects that focused on VOT and other secondary cues also showed that word-medial voiceless plosives might be fully or partially neutralized with voiced plosives in terms of VOT [4, 9]. [4] investigated the VOT of word-initial and word-medial plosives and found no significant difference between voiced and voiceless plosives in the word-medial environment. In our follow-up study [9], VOT measurement was refined to include fully pre-voiced tokens and confirmed partial neutralization in VOT; however, other acoustic cues, including F0 and duration, do not seem to play an active role in voicing contrast realization. The phonological voicing and VOT polarity of both dialects are summarized in Table 1.

	Word Initial	Word Medial
Tokyo	Voiced: −/+ Voiceless: +	Voiced: − Voiceless: +
Tohoku	Voiced: −/+ Voiceless: +	Voiced: − Voiceless: −/+

Table 1: Intervocalic voicing and initial devoicing in the two dialects (−, +: VOT polarity).

The following limitations of the previous studies can be addressed: (1) Most previous research data are based on spontaneous speech, which enhances the difficulty of evaluating any possible effect resulting from the phonological environments, including following vowels, accentual types, and lexical strata, primarily when the data set includes F0. (2) A wide

variety of Tohoku dialects were targeted, which could be an issue since each dialect may have a different extent of intervocalic voicing. Let alone gender and age were also reported as important factors of intervocalic voicing [10, 11]. An experiment is necessary to assess the above factors appropriately. (3) The influence of Tokyo Japanese is challenging to evaluate under the flow of language standardization. A framework within which the realization of both Tokyo and Tohoku dialects can be adequately evaluated is needed.

To investigate the role of VOT and other acoustic cues in the realization of plosives in the Tokyo and Tohoku dialects of Japanese, we conducted a production experiment using a psycholinguistic factorial design that controlled for accentual types and lexical strata. Given the diversity within Tohoku dialects, only speakers from the Yamagata prefecture were included in the experiment.

2. METHODS

2.1. Experiment

To examine the realization of VOT and other acoustic cues in the plosive voicing contrast in two groups of Japanese speakers, we conducted an entirely remote production experiment using *jsPsych* [12], through *Cognition.run* (https://cognition.run). All the participants were recruited online via a web recruiting service (*Crowdworks*; https://crowdworks.jp) and completed the experiment remotely using their own computers with microphones and either speakers or headphones. To control for the potential influence of other dialects, especially the standard Japanese dialect, on Tohoku speakers, tutorials were conducted in the dialects specific to each group. The tutorials were recorded by a native male speaker from Yamagata and a native female speaker from Tokyo.

2.1.1. Participants

The participants included 10 Yamagata-dialect speakers who were born and raised in Yamagata Prefecture, and 20 Tokyo-dialect speakers, born and raised in the Tokyo-Metropolitan area including Tokyo, Saitama, Chiba, and Kanagawa. All participants completed a consent form before beginning the experiment. The data of one speaker from Yamagata and two speakers from the Tokyo-Metropolitan area with incomplete experiments were excluded from the analysis. As a result, the data of 9 Yamagata-dialect speakers (female: 4, male: 5; average age: 38.44 years, *SD* of age: 9.89) and 18 Tokyo-dialect speakers (female: 7, male: 11; average age: 37.22 years; *SD* of age: 10.60) were analyzed.

2.1.2. Experimental design

This experiment contains two within- and one between-subject factor, all of each with two levels as follows: *voicing* (voiced/voiceless), *word position* (initial/medial), and *group* (Tohoku/Tokyo). Four conditions were created as shown in Table 2.

	Voiced	Voiceless
Initial	***g**oma* "sesami"	***k**izu* "scar"
Medial	*fu**d**e* "pen"	*ha**t**o* "pigeon"

Table 2: Examples in each condition. Italic bold characters indicate the target consonants. Bilabial stops have been excluded due to the distinctive phonological contrast present in Japanese ([h, b] rather than [p, b]).

Each within-subject condition contained 16 words with identical numbers of target segments and vowels. A total of 64 items were used in the experiment. Words of foreign stratum and those with clear morphological boundaries were not used due to the possible effects on intervocalic voicing [13]. Items were presented in the form of pictures. Each item consisted of a picture of an older adult man attempting to identify an object, followed by the target word and a carrier sentence in either the Tokyo or Yamagata dialect, as shown in Table 3. Both mean "It should be __. Yeah, it is __." Participants were asked to "tell the senior man what it is" by using the provided carrier sentence after seeing the target word written in Japanese characters (*kanji* and *hiragana*), and a recording session followed. Before submission, each participant checked their recording files. Each sentence contained two repetitions of the target word, and the order of items was fully randomized.

Tokyo group	*... dayo. Sōsō, ... da.*
Tohoku group	*... dabe. Ndanda, ...da.*

Table 3: Carrier sentences used in the experiment.

2.2. Data analysis

Experimental trials were retrieved from *.json* produced on *Cognition.run*. All recorded files were converted from *base64* to a *.wav* format, and we applied auto-segmentation to the recorded files with *Julius* [14]. Five trained phoneticians inspected and, if necessary, corrected the segmentation boundaries. Acoustic properties were measured by *Praat* [15].

VOT values were measured as a duration between the burst and the voicing onset of the following vowel. Fully pre-voiced plosives were annotated following the method proposed in [16] and applied in [9]. F0 values for the vowel following the target plosive were measured. The first F0 detected by *Praat* was used as an onset F0 for the statistical analysis. Time-normalized F0 values measured at each 1/10 point of the entire vowel were obtained for the contour analysis. We corrected apparent F0 measurement

errors by adjusting the pitch settings of *Praat*. F0 values were converted from hertz to semitones relative to each speaker's mean F0. We also measured the duration of preceding vowels and excluded all segment labels overlapping silent pauses.

Three statistical models were applied to evaluate the effect of each acoustic cue, including VOT, F0, and duration. Not only the VOT but the duration was scaled by each participant to avoid any possible effects of speech rate. The statistical analysis was carried out with the Linear Mixed-Effects Model (LME), using *lme4* [17] and *lmerTest* [18] on *RStudio* (*R*: 4.1.2 [19]; *RStudio*: 2022.07.2 [20]). Pairwise comparisons were conducted using *emmeans* [21]. *Word position* was excluded in the statistical model for vowel duration since there was no preceding vowel in word-initial conditions. Factor coding is shown in Table 4.

	Voicing	*Word Pos.*	*Group*
0	Voiced	Initial	Tohoku
1	Voiceless	Medial	Tokyo

Table 4: Factor coding in the LME models.

Other predictors used as fixed factors in the statistical model were *age* (coded as numeric), *sex* (coded as factor), and *accent* (whether the accent nucleus is on the syllable of the target segment or not). Random factors include the intercepts of *participant, item, sentence position* (the first or second reading), and *segment duration* related to speech rate, which may affect the realization of intervocalic voicing [22, 23]. In the model evaluating the effect of F0 and the duration of the following vowel, *vowel type* was added as a random factor. The optimal models were determined using backward selection.

3. RESULTS

3.1. Tohoku and Tokyo speakers' VOT differences

Table 5 and Figure 1 show the results of each condition. A significant three-way interaction of *voicing*, *word position*, and *group* indicates that the interaction between voicing and position varies in the two speaker groups.

	Estimates	*SE*	*t value*	*Pr(>|t|)*
(Intercept)	−.193	.305	−.633	.532
voicing (voi)	1.210	.084	14.425	<.001
wordpos	−1.020	.086	−11.796	<.001
group	−.302	.080	−3.754	<.001
age	.003	.007	.396	.696
sex	.150	.132	1.137	.267
accent	−.004	.049	−.084	.933
voi:wordpos	−.487	.120	−4.055	<.001
voi:group	−.023	.065	−.352	.724
wordpos:group	.374	.068	5.543	<.001
voi:wordpos:group	.461	.094	4.910	<.001

Table 5: The LME results where VOTs are normalized.

In a sub-analysis, Tohoku speakers' VOT was significantly shorter in the word-medial voiceless condition than in the initial condition (β: 1.507, *SE*: .085, *z ratio*: 17.697, $p < .001$). However, another pairwise comparison suggested that the VOT was still significantly longer than the voiced medial condition (β: −.722, *SE*: .086, *z ratio*: −8.405, $p < .001$), indicating incomplete intervocalic neutralization. No significant difference was detected in age, sex, and accent in the VOT duration of each condition. There was, however, a significant difference between the Tohoku and Tokyo groups in the initial devoicing of voiced consonants (β: .325, *SE*: .080, *z ratio*: 4.030, $p < .005$). The extent of non-negative VOT values observed in the Tohoku group was even greater than in the Tokyo group.

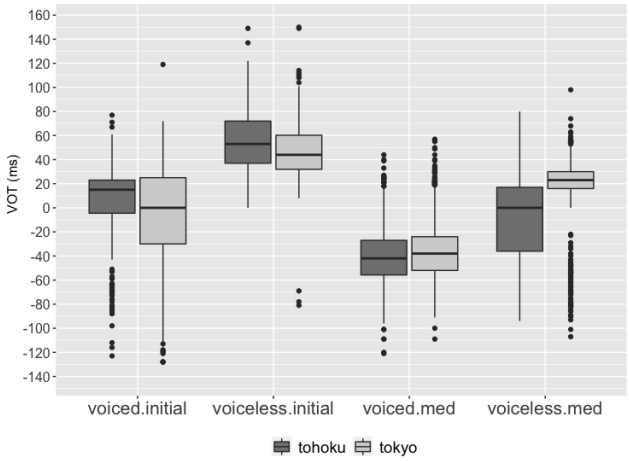

Figure 1: VOT (ms) of each experimental condition.

3.2. Results of F0 and VOT

Table 6 shows the result of F0. The three-way interaction of *voicing*, *position*, and *group* was also observed. The interaction of F0 with two within-subject factors differed in the speaker group.

	Estimates	*SE*	*t value*	*Pr(>|t|)*
(Intercept)	−2.215	1.078	−2.054	.156
voicing (voi)	2.818	.541	5.208	<.001
wordpos	2.331	.543	4.294	<.001
group	.005	.221	.024	.980
age	.002	.011	.143	.887
sex	.102	.210	.875	.390
accent	.695	.282	2.460	.015
voi:wordpos	−2.119	.765	−2.769	.008
voi:group	−.348	.245	−1.421	.155
wordpos:group	.123	.229	.543	.587
voi:wordpos:group	1.311	.332	3.951	<.001

Table 6: The LME results of the onset F0.

Further pairwise comparisons indicated that the initial voiced condition in both groups triggered a lower F0 than the voiceless initial condition ($p < .001$).

As previous studies suggested, voicing contrast in both groups involved a significant F0 change.

In the Tohoku group, voiceless medial consonants provoked a significantly lower onset F0 than voiceless initial consonants (p = .002) but did not significantly differ from voiced medial consonants (p = 1.000) in the onset region, followed by a steady descending by time as shown in Figure 2.

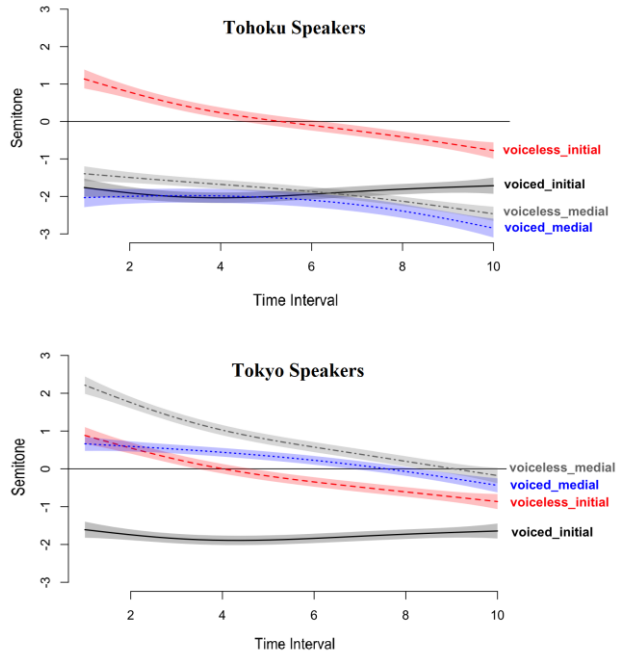

Figure 2: Plots of VOT and F0 in semitone of both speaker groups in which smooths are estimated from a fitted GAMM model using *mgcv* [24].

3.3. Durational cue of the preceding vowel

The duration result revealed a significant interaction between *voicing* and *group* (β: −.302, *SE*: .087, *t value*: −3.478, $p < .001$). The effects of *age*, *sex*, and *accent* were all of no significance. In a pairwise comparison, the duration of the preceding vowel before the voiceless plosives was slightly shorter than voiced conditions in both speaker groups (Tokyo: $p < .001$; Tohoku: $p < .001$).

4. DISCUSSION

The results confirmed the intervocalic voicing neutralization in the Tohoku dialect, which has been reported by previous studies like [4]; however, the neutralization is incomplete, which supports prior analyses [9]. The devoicing of word-initial voiced plosives was observed in both the Tokyo and Tohoku dialects. Positive VOTs in the initial voiced segments were more pronounced in Tohoku Japanese than in Tokyo Japanese, which is consistent with the results of a previous study [6] where the Yamagata dialect was categorized as a "coastal dialect of the Sea of Japan" with a higher frequency of positive VOTs among Tohoku dialects. However, no generational differences were observed in our study.

Incomplete neutralization may leave some other utilizable phonetic cues in speech [25, 26]. Following [9], this study measured the onset F0 of the following vowel and the duration of the preceding short vowels. Despite broad overlapping, the onset F0 showed a difference between the initial voiced and voiceless condition in the Tokyo group, which successfully duplicated the result in previous studies, and between the medial voiceless and voiced condition. Tokyo speakers may realize a higher onset F0 after voiceless plosives, however, in the Tohoku group, no significance between voiced and voiceless medial conditions was observed.

Durational cues also seemed to be realized, another important finding of this study. Similar to earlier studies on English [27, 28], vowels preceding voiceless segments were shorter than those preceding voiced segments in Japanese. Note that the voicing here is phonological following the experimental design. Although intervocalic voicing and initial devoicing changed the VOTs of the segments, both groups show a clear tendency for duration change.

Future research should address incomplete neutralization issues by conducting perception experiments to determine whether the identified secondary cues are also involved in the perception.

5. CONCLUSION

In this study, we utilized an online production experiment to examine intervocalic voicing and initial devoicing in Tohoku and Tokyo Japanese. The results empirically demonstrated that intervocalic voicing neutralization is incomplete in Tohoku Japanese, as evidenced by the VOT distribution. Additionally, non-negative VOTs were observed in the initial voiced conditions in both dialects. The preceding vowel duration was also found to vary with voicing in both dialects, whereas the onset F0 was found to be distinguished only in Tokyo Japanese. These findings are consistent with previous studies, which have demonstrated that vowels preceding voiceless segments in both Japanese dialects are significantly shorter than those preceding voiced segments, despite the phonological contrast of vowel length.

6. ACKNOWLEDGMENT

This research was supported by JSPS Kakenhi 22K00516.

7. REFERENCES

[1] H. Inoue, "Consonant system of the Tohoku-dialect," *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, vol. 1968, no. 52, pp. 80–98, 1968.

[2] J. Ohashi, *Tohokuhogen onsei-no-kenkyu*. Ohfu, 2002.

[3] A. Hashimoto, "Tohoku-hogen-ni-okeru kagyo tagyo-no-yuseion-ni-tsuite: Tohokuhogen onseichosa-kara (2)," *Tsuda Journal of Language and Culture*, no. 34, pp. 88–102, 2019.

[4] A. Mizoguchi, A. Hashimoto, S. Matsui, S. Imatomi, R. Kobayashi, and M. Kitahara, "Neutralization of Voicing Distinction of Stops in Tohoku Dialects of Japanese: Field Work and Acoustic Measurements," *Interspeech 2020*, pp. 1873–1877, 2020.

[5] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[6] M. Takada, *Nihongo-no gotō-heisaon-no kenkyū: VOT-no kyōjiteki-bunpu-to tsūjiteki-henka [A study of Japanese word-initial glottal stops: synchronic distribution of VOT and diachronic changes]*. Kuroshio Publishers, 2011.

[7] J. Gao and T. Arai, "Plosive (de-) voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery," *Journal of Phonetics*, vol. 77, p. 100932, 2019.

[8] H.-G. Byun, "Acoustic characteristics for Japanese stops in word-initial position: VOT and post-stop f0," *Journal of the Phonetic Society of Japan*, vol. 25, pp. 41–63, 2021.

[9] H. Noguchi *et al.*, "VOT and F0 perturbations for the realization of voicing contrast in Tohoku Japanese," in *Interspeech 2022*, ISCA, Sep. 2022, pp. 3428–3432. doi: 10.21437/Interspeech.2022-587.

[10] M. Takada, "Regional and generational variation of VOT in Japanese word-initial stops," *Papers from the First International Conference on Asian Geolinguistics*, pp. 273–282, 2012.

[11] G. Hiratsuka, "About the Voicing Phenomenon of ka, ta line Consonants Appearing in the Middle and End of Words in the Northern Part of Akita Prefecture," *Shigen : Tokyo University of Foreign Studies Descriptive Linguistic papers*, no. 13, pp. 151–158, 2017.

[12] J. R. De Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behavior research methods*, vol. 47, no. 1, pp. 1–12, 2015.

[13] N. Kibe, "Kagoshima-oyobi-Tohokuhogen-no gochu-kagyo-tagyo-no shiin-ni tsuite," 1990.

[14] T. Kawahara, "Open-source speech recognition software Julius," *JSAI*, vol. 20, no. 1, pp. 41–49, 2005.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer programme], version 6.1. 41." 2021.

[16] A. S. Abramson and D. H. Whalen, "Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of phonetics*, vol. 63, pp. 75–86, 2017.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," *Journal of statistical software*, vol. 82, pp. 1–26, 2017.

[19] R Core Team, "R: a language and environment for statistical computing. Version 4.1.2," *R Foundation for Statistical Computing, Vienna, Austria. Freely available at https://www. r-project. org*, 2021.

[20] RStudio Team, "RStudio: integrated development for R," *Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com*, 2020.

[21] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Package 'emmeans.'" 2019.

[22] J. Magloire and K. P. Green, "A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish," *Phonetica*, vol. 56, no. 3–4, pp. 158–185, 1999.

[23] R. Puggaard-Rode, C. S. Horslund, and H. Jørgensen, "The rarity of intervocalic voicing of stops in Danish spontaneous speech," *Laboratory Phonology*, vol. 13, no. 1, 2022.

[24] S. N. Wood, "Generalized Additive Models: An Introduction with R.,(Chapman and Hall: CRC Press, Boca Raton, FL.)," 2006.

[25] N. Warner, A. Jongman, J. Sereno, and R. Kemps, "Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch," *Journal of phonetics*, vol. 32, no. 2, pp. 251–276, 2004.

[26] C. Farrington, "Incomplete neutralization in African American English: The case of final consonant voicing," *Language Variation and Change*, vol. 30, no. 3, pp. 361–383, 2018.

[27] A. S. House, "On vowel duration in English," *The Journal of the Acoustical Society of America*, vol. 33, no. 9, pp. 1174–1178, 1961.

[28] L. J. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *The Journal of the Acoustical Society of America*, vol. 51, no. 4B, pp. 1296–1303, 1972.