INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Perception of incomplete voicing neutralization of obstruents in Tohoku Japanese

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

[1]Sophia University
[2]The University of Tokyo
[3]Tokyo Medical and Dental University
[4]Nagoya Gakuin University
[5]Tokyo Kasei-gakuin College
[6]Maebashi Institute of Technology, [7]NINJAL

mafuyu@sophia.ac.jp, n_watabe@phiz.c.u-tokyo.ac.jp, noguchi425@gmail.com, huang@ngu.ac.jp, hassy@kasei-gakuin.ac.jp, aimizoguchi@maebashi-it.ac.jp

Abstract

Intervocalic voicing neutralization has been generally accepted as a peculiar feature of Tohoku dialects. The present paper reports the results of perception experiments on this phenomenon. Natural and resynthesized stimuli spoken by Tohoku speakers were presented to both Tohoku and Tokyo listeners in a series of online experiments. A comparison between these two listener groups reveals that, for Tohoku listeners whose perception was biased by the voicing neutralization in their phonology, the boundary between voiced and voiceless tokens was more blurred compared to Tokyo listeners whose phonology had no such neutralization. These results suggest that neutralization can be bidirectional: i.e., voiced tokens become less voiced and voiceless tokens become less voiceless in contrast to the traditional view of neutralization which assumes a unidirectional process where one category remains intact and the other category merges with the former.

Index Terms: intervocalic voicing, VOT, neutralization, dialectal variation,

1. Introduction

Intervocalic voicing neutralization observed in Tohoku dialects of Japanese has garnered considerable attention in previous studies. While some studies have taken an impressionistic or descriptive approach [1]–[3], others have examined the issue through acoustic or quantitative studies [4], [5]. The results of acoustic measurements suggest that an incomplete neutralization pattern exists in the Tohoku dialects. In particular, the voice onset time (VOT) distribution for voiceless obstruents in the word-medial position peaked at approximately 30ms, whereas that for a voiced obstruent was around -60ms, with a significant degree of overlap between the two peaks. This overlap resulted in some words, such as *kaki* "persimmon", being realized as *kagi* "key" by certain speakers. Notably, a considerable degree of inter- and intra-speaker variability was observed in the production data. Furthermore, the study found that neither the onset F0 nor the duration of the following vowel provide sufficient cues to distinguish the voicing of the obstruent in the overlapped region even though they covary with VOT.

A natural question, then, is whether and how listeners perceive the voicing distinction in the overlapped region where incomplete neutralization occurs. Perception of incompletely neutralized segments has been an intriguing research topic as an intersection of phonology, phonetics, and psycholinguistics. In particular, word-final obstruents have been extensively investigated in the literature, such as German [6], [7], Dutch [8], [9], and Polish [10] among others. The term "incomplete neutralization" tacitly assumes a *complete* neutralization, at least at the broad level of transcription. The rationale behind the term is that so-called phonologically neutralized segments are *not completely* the same because speakers produce them slightly differently and listeners can use such subtle phonetic cues to distinguish them. Some production studies have found differences in the duration of the preceding vowel, closure, closure voicing, and burst [7], [10]. However, other production and perception studies suggested that speech style, context, orthography, and distribution of morphemes in the lexicon may play some roles in making such subtle differences in production and biased responses in perception [8], [9].

In Tohoku dialects, the phenomenon of voicing neutralization appears to be less prominent, with a considerable proportion of tokens exhibiting discernible differences. Nevertheless, previous descriptive studies maintain that voicing neutralization represents a regular and persistent trait of Tohoku dialects [2], [11]. We must reconcile the observed discrepancies between phonological and phonetic perspectives. Specifically, an investigation of the perception of the overlapped region in a VOT continuum holds significance. Moreover, comparing Tohoku listeners' discrimination of stimuli with those of the Tokyo dialect, known for its non-neutralizing characteristics at the intervocalic position, offers a novel insight into this issue.

Given the above considerations, the following research questions have been formulated:

(1) Do Tohoku listeners accurately distinguish minimal pairs of natural tokens that contrast in the voicing of an intervocalic obstruent?

(2) What are the cues and where is the boundary for Tohoku listeners to distinguish minimal pairs in voicing?

10.21437/Interspeech.2023-2103

(3) How do Tokyo listeners respond to the same sets of stimuli in (1) and (2)?

2. Methods

2.1. Participants

In this study, we conducted three consecutive online perception experiments, utilizing jsPsych [12] on Cognition.run (https://www.cognition.run/), to investigate whether these parameters function as secondary cues. Participants were primarily recruited through the CrowdWorks web platform (https://crowdworks.jp/) and completed the experiments on their personal computers. The study involved 77 participants (*Mean* of age: 38.26; *SD*: 10.46), including 40 native speakers of Tokyo Japanese and 37 native speakers of Tohoku Japanese, all of whom provided informed consent prior to their participation.

2.2. Experiment 1

The initial experiment employed recordings obtained from a production study, where the carrier sentences described in (1) were utilized. The experiment involved the extraction of tokens that comprised the target segments, along with the subsequent particles, articulated by nine (4 females and 5 males) speakers of Tohoku Japanese. Each word was repeated twice. The items are detailed in (2).

(1) ... dabe. Ndanda, ...da. (It is …, right, … is it.)

(2) a. kaki (persimmon)/kagi(key)
b. mato(target)/mado(window)
c. kuki(stem)/kugi(nail)

The total number of tokens was 108 (6 words * 2 repetitions * 9 speakers). The extracted sounds were played in a randomized order, and participants were presented with two buttons each containing pictures of the words with voiceless target segments or their voiced counterparts. They were then instructed to select the picture that best matched the sound they heard.

2.3. Experiment 2

In this experiment, we manipulated sounds using Praat [13]. We selected one pair of words with identical pitch patterns in Tokyo Japanese (2a) *kaki* – *kagi* spoken by three different speakers. Using a Praat script, we created an 11-step VOT continuum for each token by shortening the VOT of the second consonant in *kaki* and compensating for the total duration by lengthening the surrounding vowels. For example, one of the three original tokens showed 81ms of VOT. The decrement step for this token was 1/10 of the original VOT, i.e., 8.1ms. Eleven stimuli were created from Level 0 (81ms) to Level 10 (0ms) in even intervals.

2.4. Experiment 3

The method in Experiment 2 only allowed us to modify the absolute values of VOT within the same polarity. In other words, only *kaki* was used as the original material to create VOT-clipped tokens. Some potentially relevant acoustic properties included in the original *kagi* could not be reflected in the stimulus set. To overcome this limitation, a vocoder-based analysis and resynthesis platform, WORLD [14], [15] was employed in Experiment 3. In addition, a morphing generation procedure [16] was utilized to create the continuum.

The same three *kaki* tokens as in Experiment 2 were used as the original materials, with corresponding *kagi* tokens included in the resynthesis process: (1) the pair produced by one speaker was time-aligned with respect to the burst position and vowel portions; (2) *kaki* token was set to one end and *kagi* token was set to the other end of the morphing continuum; (3) time-axis morphing rate and the spectrum-level morphing rate were set to generate 9 equidistant steps between the two ends; (4) the same steps were repeated for two more speakers' tokens. There were 3 sets of 11-stimuli continua as the outcome.

3. Results

3.1. Experiment 1

In the first experiment, responses corresponding to the phonological representation which Tohoku speakers aimed for were coded as 1. Generalized Linear-Mixed Models (GLMM; package lme4 [17]) were used for analysis. Experimental factors were contrast-coded as fixed factors in the model, along with their interaction: voicing (voiced = 0/voiceless = 1) and group (Tokyo = 0/Tohoku = 1). Other fixed factors included gender (female = 0; male = 1) and age (as a numeral). Random factors included the intercepts of the item and the participant. The optimal model was selected through backward selection. The results are presented in Table 1, Figure 1, and Figure 2.

Table 1. *Correct response rates for voiced and voiceless items by Tokyo and Tohoku listeners.*

Voicing	*Group*	*Mean*	*SD*
voiced	Tokyo	0.922	0.066
voiced	Tohoku	0.814	0.103
voiceless	Tokyo	0.502	0.077
voiceless	Tohoku	0.584	0.116

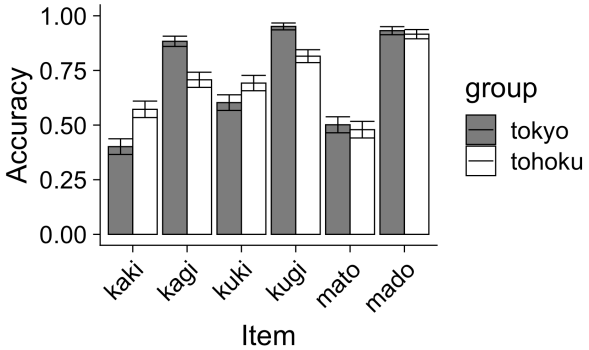

Figure 1. *Accuracy of each stimulus item by Tokyo and Tohoku listeners*

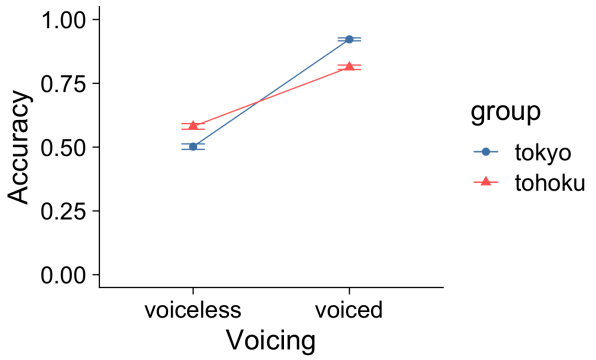

Figure 2. *Correct response rates for voiced and voiceless items by Tokyo and Tohoku listeners. Error bars indicate the standard errors of the mean.*

The statistical model revealed that, overall, Tohoku voiced segments were perceived more accurately than voiceless segments (*Estimate*: -2.799 *SE*: 0.353, *z-value*: -7.931, $p < .001$). The accuracy of Tohoku listeners was significantly lower than that of the Tokyo group (*Estimate*: -1.101, *SE*: 0.380, *z-value*: -2.899, $p < .001$). A significant interaction was observed between voicing and group. Post-hoc analysis using the Bonferroni method on *emmeans* [18] revealed that the accuracy of the voiced condition in the Tokyo group was higher than that of the Tohoku group. Age and gender did not show any significant differences.

Table 2. *Statistical analysis of Experiment 1 using GLMM*

	Estimate	*Std.Error*	*z-value*	*Pr*
(Intercept)	2.867	0.314	9.140	< .001
voicing	-2.799	0.353	-7.931	< .001
group	-1.101	0.380	-2.899	< .001
age	0.001	0.003	-0.101	.920
gender	-0.087	0.071	-1.219	.223
voicing: group	1.425	0.452	3.152	< .001

3.2. Experiment 2

In Experiment 2, we investigated whether there were differences in voicing perception between the Tokyo and Tohoku groups of Japanese listeners. To test this, we used a VOT continuum with varying degrees of voicing, ranging from completely voiceless to fully voiced. We analyzed the responses using Generalized Linear-Mixed Models (GLMM; package lme4), including fixed factors of group (Tokyo = 0/Tohoku = 1) and VOT continuum, which was dummy-coded. Level 0, representing sounds without manipulation, was used as the baseline, and other levels were compared to this baseline. Gender and age were also included as fixed factors, while the same random factors as in the previous model were included.

Figure 3 shows the results of the analysis, with the y-axis indicating the rate of perceiving segments as voiced. As the VOT of the target segment was further apart from Level 0, both groups increasingly perceived it as voiced. However, a significant difference emerged between the two groups from Level 8 of the VOT continuum, with the difference remaining significant at Levels 9 and 10 (Level 8: *Estimate*: -1.007, *SE*: 0.331, *z-value*: -3.045, $p < .005$.Level 9: *Estimate*: -0.801, *SE*: 0.337, *z-value*: -2.374, $p < .05$. Level 10: *Estimate*: -0.890, *SE*: 0.2501, *z-value*: 11.040, $p < .001$). In addition, male participants perceived more stimuli as voiceless than female participants (*Estimate*: -0.269, *SE*: 0.135, *z-value*: -1.994, $p < .005$). The optimal model was selected through backward selection. The selection of the parameters was the same as that in Experiment 1 except for the item because only one pair was manipulated.

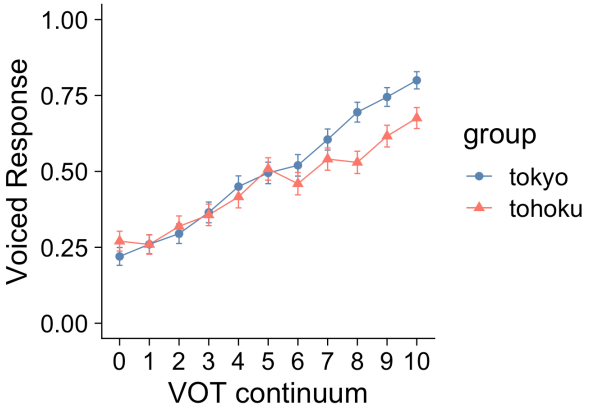

Figure 3. *Response rates of Tokyo and Tohoku listeners as a function of VOT continuum (0 for unmanipulated, 10 for zero VOT). Error bars indicate the standard errors of the mean.*

The results suggest that when the VOT is short and close to zero, Tokyo Japanese listeners are more likely than Tohoku listeners to perceive the segment as voiced. This finding is consistent with previous research that has demonstrated differences in the perception of voicing across different dialects and regions.

3.3. Experiment 3

To further investigate the differences in voicing perception between the Tokyo and Tohoku groups of Japanese listeners, we conducted Experiment 3, in which we used a set of stimulus continua with different degrees of voicing created by the morphing method. In this experiment, Level 5 of the continuum was coded as the baseline following a dummy-coding method, and the responses were analyzed using GLMM, with group (Tokyo = 0/Tohoku = 1) and the morphing continuum contrast-coded as fixed factors. Additionally, we examined whether the

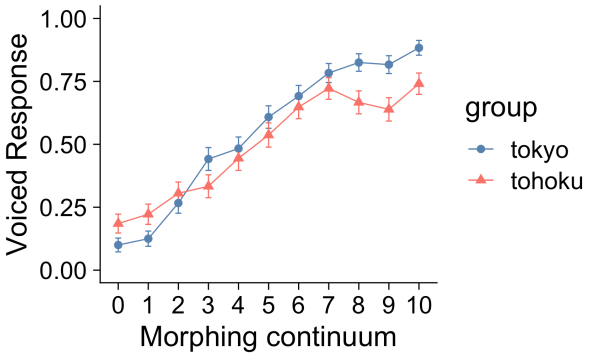

Figure 4. *Response rates of Tokyo and Tohoku listeners as a function of the morphing continuum. Error bars indicate the standard errors of the mean.*

Tokyo and Tohoku groups differed not only in their perception of voiceless segments but also in their perception of voiced segments.

Our analysis, as shown in Figure 4, revealed that both groups had a similar pattern of voicing perception as in Experiment 2 in the levels with larger numbers. However, we also found that the Tohoku group perceived Level 0 and Level 1 as voiceless segments, which were significantly lower than the Tokyo group. (Level 0: Estimate: 0.984, SE: 0.489, z-value: 2.011, $p < .05$; Level 1: Estimate: 0.977, SE: 0.463, z-value: 2.109, $p < .05$). The results of Experiment 3 suggest that the perceptual difference in voiced segments may be related to the differences of phonological voicing characteristics between the two dialectal variants of Japanese.

4. Discussion

Research question (1) raised in the introduction section was the following:

(1) Do Tohoku listeners accurately distinguish minimal pairs of natural tokens that contrast in the voicing of an intervocalic obstruent?

Results in Experiment 1 suggest that the performance of Tohoku listeners was far from accurate: underlyingly voiced tokens were identified correctly 81.4% of the time, and underlyingly voiceless tokens were identified much more poorly (58.4%). As expected from the overlap of VOT in the stimuli, many of the tokens are ambiguous and phonetic cues alone are insufficient to identify them correctly.

However, Tokyo listeners responded to the same set of stimuli in an interestingly different manner as depicted in Figure 2. A significant group-by-voicing interaction and post-hoc analyses indicate that Tokyo listeners were better at identifying underlyingly voiced tokens than Tohoku listeners, which was reversed for underlyingly voiceless tokens. In other words, voicing neutralization affected Tohoku listeners more negatively than Tokyo listeners for the underlyingly voiced tokens, and vice versa for underlyingly voiceless tokens. We can interpret this as follows: For Tohoku listeners whose perception was biased by the intervocalic voicing neutralization in their phonology, the boundary between voiced and voiceless tokens was more blurred compared to Tokyo listeners whose phonology had no such neutralization in that position. Thus, Tokyo listeners could take the face value of the voiced tokens and identify them as voiced more accurately than Tohoku listeners. Tokyo listeners became confused with underlyingly voiceless tokens pronounced with unfamiliar VOT values to them and thus their performance was close to the chance level.

To further investigate the nature of the neutralization from the perceptual side, we raised the second research question:

(2) What are the cues and where is the boundary for Tohoku listeners to distinguish minimal pairs in voicing?

In Experiment 2, VOT values of phonetically voiceless tokens were manipulated to create a continuum. Level 0 in the continuum was the unmanipulated voiceless token with a positive VOT value. Both Tohoku and Tokyo listeners showed similar response rates from Level 0 up to Level 5 at which the response rate reached about 50%. In Levels 8-10, the two groups diverged in the response rate, which was statistically significant. As shown in Figure 3, Tokyo listeners tend to judge those stimuli as voiced more than Tohoku listeners. This tendency is partially consistent with the results in Experiment 1 where Tokyo listeners more accurately perceived the unmanipulated voiced tokens. VOT continuums created by resynthesis reveal that the 5th step (35.8 ms in VOT on average) is the boundary for categorization in both groups. The voicing neutralization in Tohoku phonology affected the perception by the Tohoku group in the short VOT range, however. They perceived VOT-clipped *kaki* as more ambiguous between *kaki* and *kagi* compared to the Tokyo group.

In Experiment 3, the stimuli were constructed by a morphing method bridging between voiceless *kaki* and voiced *kagi*. The midpoint stimulus, Level 5 was to be the most ambiguous one for listeners, while Level 0 and Level 10 were the least ambiguous. As shown in Figure 4, the results conformed to the line of interpretation for Experiment 1, and the morphing method improved the situation in Experiment 2. Both Tohoku and Tokyo listeners responded to the voiced end of the stimuli as voiced, and to the voiceless end of the stimuli as voiceless. However, the difference is statistically significant on both ends: at Levels 0-1 and 8-10, Tokyo listeners are more sensitive to the acoustic differences in the continuum than Tohoku listeners, which supports our interpretation that Tohoku listeners' phonological grammar mediated their perception. Here, the difference in the sensitivity of the two groups is more clearly represented in the voiceless end than in Experiment 2.

Answers to our third research question,

(3) How do Tokyo listeners respond to the same sets of stimuli in (1) and (2)?

are now clear. Tokyo listeners, who do not have voicing neutralization as a part of their phonology, respond to both natural and resynthesized stimuli more distinctly than Tohoku listeners do.

5. Conclusions

The present paper reports the results of perception experiments on voicing neutralization in Tohoku dialects. Natural and resynthesized stimuli spoken by Tohoku speakers were presented to both Tohoku and Tokyo listeners. Comparing these two listener groups offers novel insights into both dialectal phonology and the nature of incomplete neutralization.

As expected from the production data, Tohoku listeners did not easily discern the overlapped region of VOT. Phonological neutralization in the dialect blurs the voicing distinction and phonetic cues are insufficient to distinguish ambiguous tokens. Tokyo listeners, as a control, generally performed better in the identification task.

As to the nature of incomplete neutralization, the case of the Tohoku dialect suggests that neutralization can be bidirectional, i.e., voiced tokens become less voiced and voiceless tokens become less voiceless. The traditional view of neutralization, on the contrary, assumes a unidirectional process: only voiced segments become voiceless as in the cases of German, Dutch, and Polish. The term incomplete neutralization, then, claims that underlyingly voiced segments do not completely merge with voiceless segments.

6. Acknowledgements

This research was supported by JSPS KAKENHI Grant Number JP22K00516.

7. References

[1] N. Tsujimura, *An Introduction to Japanese Linguistics*. Cambridge, MA: Blackwell, 1996.

[2] J. Ohashi, *Tohoku hougen onsei no kenkyuu [Research on sounds of Tohoku dialects]*. Tokyo: Oufuu, 2002.

[3] M. Shibatani, *The Languages of Japan*. Cambridge, UK: Cambridge University Press, 1990.

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

[5] 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," in *Interspeech 2020*, Oct. 2020, pp. 1873–1877. doi: 10.21437/Interspeech.2020-3191.

[6] M. Fourakis and G. K. Iverson, "On the 'incomplete neutralization' of German final obstruents," *Phonetica*, vol. 41, no. 3, pp. 140–149, 1984, doi: doi:10.1159/000261720.

[7] R. F. Port and M. L. O'Dell, "Neutralization of syllable-final voicing in German," *J. Phon.*, vol. 13, no. 4, pp. 455–471, Oct. 1985, doi: 10.1016/S0095-4470(19)30797-1.

[8] N. Warner, A. Jongman, J. Sereno, and R. Kemps, "Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch," *J. Phon.*, vol. 32, no. 2, pp. 251–276, Apr. 2004, doi: 10.1016/S0095-4470(03)00032-9.

[9] M. T. C. Ernestus and R. H. Baayen, "Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch," *Language*, vol. 79, no. 1, pp. 5–38, 2003, doi: 10.1353/lan.2003.0076.

[10] L. M. Slowiaczek and D. A. Dinnsen, "On the neutralizing status of Polish word-final devoicing," *J. Phon.*, vol. 13, no. 3, pp. 325–341, Jul. 1985, doi: 10.1016/S0095-4470(19)30763-6.

[11] Inoue, Fumio, "Tohoku hougen no shiin taikei [Consonant Sytem of the Tohoku-Dialect]," *Gengo Kenkyu*, vol. 52, pp. 80–98, 1968.

[12] J. R. de Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, Mar. 2015, doi: 10.3758/s13428-014-0458-y.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Nov. 01, 2021. Accessed: Nov. 10, 2021. [Online]. Available: http://www.praat.org/

[14] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, Nov. 2016, doi: 10.1016/j.specom.2016.09.001.

[15] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016, doi: 10.1587/transinf.2015EDP7457.

[16] Kawahara H. and Morise M., "Issues emerged from implementation of GUI tools for WORLD VOCODER," *IEICE Tech. Rep.*, vol. SP2022, no. 2, 2022.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4." arXiv, Jun. 23, 2014. Accessed: Mar. 01, 2023. [Online]. Available: http://arxiv.org/abs/1406.5823

[18] R. V. Lenth *et al.*, "Emmeans: Estimated marginal means, aka least-squares means." Jan. 17, 2023. [R package]. Available: https://github.com/rvlenth/emmeans