AGE- AND GENDER-BASED VARIATION IN THE PERCEPTION OF VOICING CONTRAST IN TOKYO JAPANESE[*]

Yoonjung Kang[1], Manami Hirayama[2]

University of Toronto Scarborough[1], Seikei University[2]
yoonjung.kang@utoronto.ca, hirayama@fh.seikei.ac.jp

ABSTRACT

Recent studies report a sound change in progress in Tokyo Japanese, whereby word-initial voiced stops are frequently devoiced, and VOT alone is no longer a sufficient or reliable cue to distinguish the voicing contrast. The current study examines how Tokyo Japanese speakers of different age and gender use VOT, as well as the following vowel's pitch (F0) and voice quality (h1-h2) in voicing perception. 140 speakers of Tokyo Japanese, balanced for age and gender, participated in an online perception experiment. The majority of speakers made use of all three cues, but among the three cues, some speakers relied relatively more on VOT, while others relied more on the vocalic cues, especially F0. We found a significant interaction of F0 cue use with pitch accent, gender, and age, whereby the dominant cue for voicing perception shifts from VOT to F0, and this change is led by accented words and younger females.

Keywords: Japanese, sound change, perception, VOT, F0

1. INTRODUCTION

Recent studies on Tokyo Japanese report age- and gender-based variation in the realization of the word-initial stop voicing contrast [1-3]. Specifically, word-initial voiced stops vary between prevoiced and devoiced realizations, and the rate of devoicing is higher for younger than older speakers [1-3] and for female than male speakers [2]. This change creates an overlap in VOT (Voice Onset Time) of the voiced and voiceless categories, and the contrast cannot be reliably distinguished by VOT alone.

Production studies show that the voicing contrast is signalled by other secondary cues, namely, the pitch and voice quality of the following vowel, as well as VOT. The pitch (F0) is higher when the vowel follows a word-initial voiceless stop than a voiced stop [4-7], and the voice quality of the vowel is breathier (i.e., higher h1-h2) for voiceless than voiced stops [5, 6].

Studies have also investigated the perceptual cues that listeners use to distinguish initial stop voicing. For instance, [8] examined the perception of a monosyllabic word pair manipulated to vary in F0 and VOT (positive VOT values only) and found that the effect of F0 is only noticeable for very short VOT values. The study involved college students from four dialect regions, including Kanto, which encompasses Tokyo. Another study ([9]) created two sets of stimuli that represented different pitch accent conditions: *pasu* 'pass' vs. *basu* 'bus' for initial accented words (henceforth #HL) and *teki* 'enemy' vs. *deki* 'result' for unaccented words (henceforth #LH). The stimuli were manipulated to vary in F0 and VOT, covering both positive and negative VOT values. They found that the effect of F0 was more pronounced for the #HL than the #LH condition, which is consistent with [7]'s finding in production that initial stop voicing has a stronger effect on F0 for H-initial than L-initial words. However, it is unclear whether the interaction of pitch accent and F0 cue use is due to the larger pitch range in the #HL stimuli or whether listeners weigh the F0 cue more in the #HL condition.

Finally, one study [1] examined the age-based variation in perception of voicing contrast, using naturally produced tokens of stops. The majority of the participants in the study were from Tokyo but included speakers from other dialect regions. The study found that errors that misidentify devoiced voiced stops as voiceless were no more frequent with older listeners than younger listeners, contrary to the expected pattern given that older speakers are less likely to devoice voiced stops than younger listeners in production. Instead, overall, more errors were found for the younger speakers and their errors were more widespread, regardless of whether the stop was prevoiced or not. However, the observation about age difference is difficult to interpret as the low count and the limited distribution of errors by older listeners may be a function of the low number of older participants.

Building on these previous studies, our study examines the perceptual cues used by Tokyo speakers to identify word-initial voiced and voiceless stops, and how cue use varies as a function of pitch accent and speakers' age and gender. Given the report that the change in Tokyo Japanese is led by younger females and that the VOT cue is weakening in the speech of younger and/or female speakers, we expect the sound change to be reflected in our data. In particular, we hypothesize that female and younger speakers will exhibit more innovative cue weighting and pay more attention to vocalic cues compared to male and older counterparts. This is based on the assumption that the speaker's own production and perception patterns are generally expected to align with each other (cf. [10]). While similar changes have been reported in other dialects, we focus on Tokyo because of a larger body of previous studies to build upon.

Regarding pitch accent, the larger F0 difference between voiced and voiceless stops in the #HL condition compared to the #LH condition presents three potential interactions between F0 and pitch accent in perception. The first possibility is that if the range of F0 variation in the perceptual stimuli is comparable across the pitch accent conditions, listeners may give equal weight to both the #HL and #LH pitch accent conditions. The second possibility is that, given the relative salience of F0 cues in the #HL condition, listeners may weigh F0 cues more for #HL words than for #LH words. The third possibility is that listeners are less sensitive to F0 cues for the #HL condition, necessitating a larger F0 difference to shift the voicing boundary for #HL words to match the large difference observed in production.

2. METHODS

The stimuli consisted of two minimal pairs, 手前 [temae] vs.出前 [demae] (unaccented, #LH) and 天使 [tenɕi] vs. 電子 [denɕi] (initial accented, #HL). The selection of coronal stop-initial words was based on the avoidance of labials, which tend to include English loans, and dorsals, which tend to exhibit less overlap in VOT between voiced and voiceless stops [2, 7]. These words are also frequently used and considered to be familiar to native speakers of (Tokyo) Japanese [11].

The stimuli words were produced by the second author, a female Tokyo Japanese speaker in her 40s, with 10 to 12 repetitions. For each word pair, four baseline tokens were generated by splicing prevoicing (for the negative VOT baselines) or aspiration (for the positive VOT baselines) with one of two base vowel tokens (one each from a voiced and a voiceless stop production).

The duration and F0 contour of the rest of the word were manipulated to match the average of all tokens for that pair, removing potential duration and pitch cues to voicing present in the natural production. The intensity of each spliced part (prevoicing, aspiration, and vowel) was also adjusted to closely match the speaker's average value.

The manipulation parameters and produced range for each acoustic dimension for each accent condition are summarized in Table 1, determined based on the stimuli talker's production. The VOT was varied in 10 steps from -60 ms to 50 ms, at 15 ms intervals for negative values and 10 ms intervals for positive values. The F0 at the following vowel onset (at 9.1% of the vowel duration) was varied in six equidistant steps from -2.5 to 2.5 in normalized semitone for each pair. h1-h2 was not directly manipulated, but baseline tokens typical for each word were chosen. The three acoustic dimensions were orthogonally varied for each word pair, creating 240 stimuli (=10 VOT steps * 6 F0 steps * 2 (h1-h2) baseline vowels * 2 pitch accent word pairs). The manipulations were conducted in Praat [12].

	Stimuli values	Produced range	
		#LH	#HL
VOT (ms)	[-60, -45, -30, -15, 0, 10, 20, 30, 40, 50]	[-64 ~ 58]	[-41 ~ 28]
F0 (st)	[-2.5, -1.5, -0.5, 0.5, 1.5, 2.5]	[-2.5 ~ 2.5]	[-4.1 ~ 4.0]
h1-h2 (dB)	#LH: [-7.1, 0.3] #HL: [-11.4, -8.5]	[-14.8 ~ 16.0]	[-18.1 ~ 2.5]

Table 1: Acoustic parameters for the perception stimuli and the ranges for the produced tokens

	60s+	50s	40s	30s	20s
Female	15	14	12	14	11
Male	14	15	16	14	15

Table 2: Age and gender breakdown of participants included in the analysis

Self-identified Tokyo Japanese speakers from Tokyo, Chiba, Saitama, or Kanagawa were recruited through an online crowdsourcing recruitment site (crowdworks.jp). A total of 172 speakers participated, out of which four were excluded for responding "no" to the question "Do you speak Tokyo-style Japanese?" Additionally, 28 speakers who answered "yes" to the same question but also listed other dialects they speak were also excluded. The breakdown of the 140 speakers included in the analysis by age and gender is presented in Table 2. The task was word identification, whereby participants heard stimuli and chose the word they heard. The full experiment, built and conducted on the Gorilla platform (gorilla.sc) [13], included informed consent, a background questionnaire, a production task, and a perception task, and it took an average of 19.5 minutes to complete. The perception experiment alone took 9.4 minutes.

For statistical analysis, we used the lme4 package [14] in R [15] to build a logistic mixed-effects regression model. The model takes the Response (voiced = 0, voiceless = 1) as the response variable and four linguistic predictors (VOT, F0, Base.vowel, Accent), two speaker-level predictors (year of birth (YOB), GENDER), and their full interactions as fixed effects predictors. To compare their relative strength, the phonetic variables, VOT (ms), F0 (st), and BASE.VOWEL (voiced = 0, voiceless =1), were transformed into z-scores to put them on a comparable scale. ACCENT and GENDER were sum-coded (#LH = -0.5, #HL = 0.5; F = -0.5, M = 0.5), while YOB was centered. We also included by-PARTICIPANT random intercept and random slope adjustments for the three phonetic predictors, which we explore in detail to examine cue use variation across listeners (cf. [16]). The formula is shown in (1) below. We used an alpha level of 0.05 for significance.

```
(1) glmer(RESPONSE ~ VOT * F0 * BASE.VOWEL
* ACCENT * YOB * GENDER + (1 + VOT + F0
+ BASE.VOWEL | PARTICIPANT),"binomial")
```

3. RESULTS

Figure 1 plots the proportion of voiceless responses by the four linguistic conditions (VOT, F0, BASE.VOWEL, and ACCENT) pooled across all participants. The plots illustrate, and the statistical test confirms that all three phonetic predictors have significant effects. Overall, a higher VOT, a higher F0, and a voiceless BASE.VOWEL all lead to more voiceless responses. We can also observe that these predictors interact, and the statistical model shows a significant four-way interaction (VOT * F0 * BASE.VOWEL * ACCENT). In other words, the effects of the predictors vary across different contexts.

Figure 2 displays the breakdown of the data by age and gender of the participants. Although the overall patterns are similar across different age and gender groups, there is some variation. The statistical analysis revealed that the six-way interaction of all predictors was significant, indicating that the effects of the linguistic and speaker-level predictors on the proportion of voiceless responses varied across different participant subgroups. See the full results at http://individual.utoronto.ca/yjkang/icphs_j_fit.pdf.

In this paper, we solely focus on the main effects of the three phonetic cues and their interaction with non-phonetic predictors. These interactions provide insights into how cue use changes depending on the pitch accent condition and the age and gender of the participant. Table 3 summarizes the coefficient estimates and only presents interactions with significant effects. The intercept coefficient estimate (not shown in the table) is 0.702. The scatterplots in Figure 3 display the distribution of individual participants' estimated slopes for the phonetic cues calculated from the model.

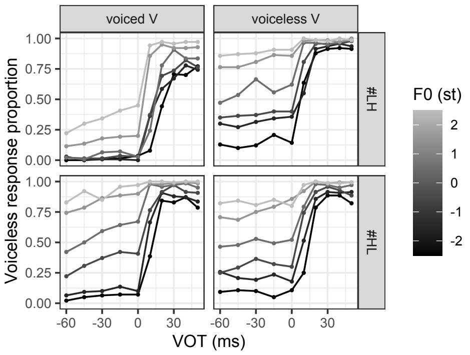

Figure 1: Proportion of voiceless responses by VOT, F0, BASE.VOWEL, and ACCENT, across all participants

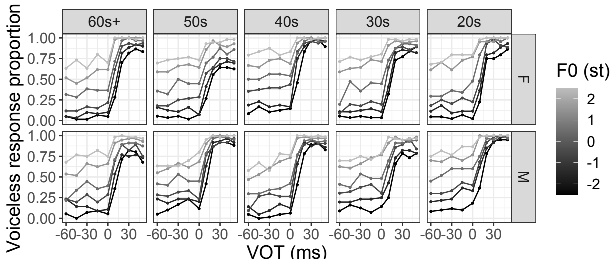

Figure 2: Proportion of voiceless responses by VOT and F0, separated by AGE and GENDER of participants.

	VOT	F0	VOWEL
main effects	2.098	1.591	0.647
2-way interaction			
* accent (#HL-#LH)	-0.573	0.326	-1.631
* gender (M-F)	(-0.054)	-0.239	0.137
* year of birth (yob)	(-0.007)	(0.053)	(0.012)
3-way interaction			
* accent * gender	(-0.165)	-0.241	(-0.011)
4-way interaction			
* acc. * gender * yob	(-0.093)	-0.193	(-0.158)

Table 3: Coefficient estimates of the main effects of the three phonetic predictors (VOT, F0, BASE.VOWEL) and interactions with non-phonetic predictors. Parentheses indicate non-significance effects.

All three main effects of the phonetic predictors are significant, but they interact significantly with ACCENT, suggesting that cue weights vary depending on the pitch accent condition. This can also be seen from the comparison of the top vs. bottom panels of Figure 3. For the #LH word pair [temae/demae], on average, VOT is a much stronger cue (β = 2.098 + (-0.573 * -0.5) = 2.385) than F0 (β = 1.428) or BASE.VOWEL (β = 1.463). Also note that in the top panels of Figure 3, all participants' coefficients are above 0, which means that everyone used all three cues in the expected direction. For the accented #HL word pair [teɕi]-[deɕi], on the other hand, the average coefficients are comparable between VOT (β = 1.812) and F0 (β = 1.754), but they differ in variability. In the bottom panels of Figure 3, individual slopes for F0 are tightly clustered around 2, while VOT coefficients are more variable, with some even falling below 0. BASE.VOWEL coefficients are distributed around 0, indicating that the cue does not have a consistent effect and that many participants use the cue in the opposite direction.

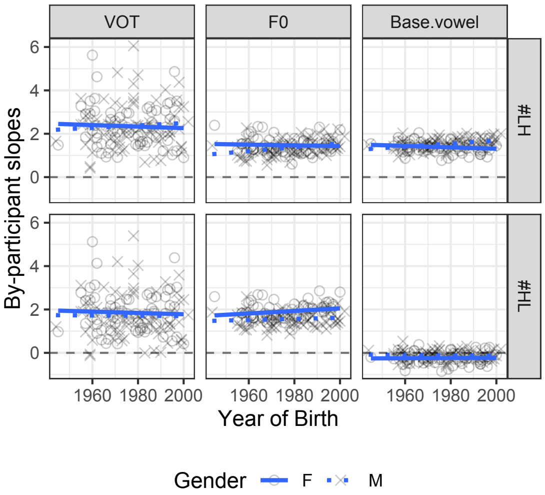

Figure 3: By-participant slope estimates for phonetic predictors, plotted by ACCENT and by the participants' GENDER and YOB. The dashed horizontal lines mark the slope of 0 and the solid and dotted lines are linear smooths for female and male participants.

Regarding the speaker-level predictors, AGE and GENDER, if individuals' perception reflects the sound change in progress where female and younger speakers are more likely to devoice voiced stops and reduce the effectiveness of the VOT cue in their speech than male and older speakers, we expect female and younger speakers to weight vocalic cues (F0 in particular) more than male and older speakers respectively, while VOT cues should show the opposite trend. This prediction holds for F0. The significant interactions of F0 * GENDER, F0 * ACCENT * GENDER, and F0 * ACCENT * GENDER * YOB show that overall, the F0 cue is stronger for females than males, and this interaction is stronger for the #HL [teɕi]~[deɕi] pair and for younger participants. On the other hand, VOT does not show significant interactions with YOB or GENDER.

Turning to individuals' use of cues, we can classify the participants according to the cue they pay the most attention to, that is, the cue with the highest coefficient. Table 4 tabulates the number of participants by their dominant cues. For example, among the female participants in their 20s (n = 10), VOT was the best cue for four, and F0 was the best cue for the other six. None of the 10 participants weighted BASE.VOWEL as the most important cue.

The age- and gender-based variation in dominant cues at the individual level mirrors the significant interactions of F0 with GENDER, AGE and ACCENT in the model. For the unaccented (#LH) pair, VOT is still the dominant cue regardless of participants' age and gender, while for the accented (#HL) pair, which overall shows raised sensitivity to F0 and reduced sensitivity to VOT, proportionally more females use F0 as the dominant cue (57.6% = 38 out of 66) than males (50.0 % = 37 out of 64). Also, note that the majority dominant cue shifts from VOT to F0 as we move from older to younger speakers. In other words, we see the cue shifting from VOT to F0 for the younger speakers, and the change is more advanced in female participants and for accented words.

	Age	Female			Male		
		VOT	F0	vowel	VOT	F0	vowel
#LH	60s+	13*	2	0	10*	2	2
	50s	9*	3	2	12*	0	3
	40s	10*	0	2	14*	1	1
	30s	11*	2	1	9*	2	3
	20s	8*	1	2	11*	2	2
#HL	60s+	8*	7	0	8*	6	0
	50s	5	9*	0	10*	5	0
	40s	4	8*	0	9*	7	0
	30s	6	8*	0	4	10*	0
	20s	4	6*	0	6	9*	0

Table 4: The distribution of participants' dominant phonetic cue based on their AGE, Gender, and the ACCENT. The asterisk (*) denotes the majority pattern for each demographic subgroup.

4. DISCUSSION

In this paper, we investigated the perceptual cue use in the context of sound change in progress for Tokyo Japanese stop voicing. Our findings suggest that pitch accent plays a significant role in the cue weighting pattern, which is consistent with the production differences observed in the stimuli talker's speech (Table 1) and reflects the speech patterns of Tokyo Japanese speakers in general [7]. Specifically, F0 differed more by voicing for the #HL pair, while VOT and BASE.VOWEL quality differed more for the #LH. We also found listeners' age and gender had an effect on perception, as expected. For the #HL pair only, younger and female listeners relied more heavily on F0 than older and male listeners. This interaction of F0 with age and gender affected the relative importance of VOT, which was relied on relatively less by younger female listeners compared to F0. Overall, our results suggest that the dominant cue for the initial voicing contrast is shifting from VOT to F0, led by #HL words and by younger female listeners.

5. REFERENCES

[1] Takada, M. 2004. VOT tendency in the initial voiced alveolar plosive /d/ in Japanese and the speakers' age. *Journal of the Phonetic Society of Japan* 8(3), 57–66.

[2] Takada, M. 2011. *Nihongo-no Gotoo Heisaon-no Kenkyuu: VOT-no Kyoojiteki Bunpu-to Tuujiteki Henka* [*Research on the word-initial stops of Japanese: Synchronic distribution and diachronic change in VOT*]. Kurosio.

[3] Takada, M., Kong, E., Yoneyama, K., Beckman, M. E. 2015. Loss of prevoicing in modern Japanese /g, d, b/. *Proceedings of the 15th ICPhS.*

[4] Shimuzu, K. 1999. A study on phonetic characteristics of voicing of stop consonants in Japanese and English. *Journal of the Phonetic Society of Japan* 3(2), 4–10.

[5] Kong, E. J., Beckman, M. E., Edwards, J. 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *Journal of Phonetics* 40(6), 725–744.

[6] Takada, M., Kong, E. J., Yoneyama, K., Beckman, M. E. 2015. Do pitch and voice quality cue word-initial "voicing" in Tôhoku Japanese? *Poster presented at the 24th Japanese/Korean Linguistics Conference.*

[7] Gao, J., Arai, T. 2019. Plosive (de-)voicing and F0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics* 77, 100932.

[8] Byun, H.-G. 2021. Perception of Japanese word-initial stops by native listeners. *Phonetics and Speech Sciences* 13(3), 53–64.

[9] Gao, J., Yun, J., Arai, T. 2019. VOT-F0 coarticulation in Japanese: Production-based or misparsing? *Proceedings of the 16th ICPhS.*

[10] Schertz, J., Clare, EJ. 2020. Phonetic cue weighting in perception and production. *WIREs Cognitive Science* 11: e1521.

[11] NTT Communication Science Laboratories. 2021. *NTT Lexicon Database: Word Familiarity (2020 resurvey and enlarged edition*). NTT Printing Corporation.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.05, retrieved 5 January 2022 from http://www.praat.org/

[13] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[14] Bates, D., Mächler M., Bolker B., Walker S., Christensen, B. R. H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, Pavel N. 2022. Linear mixed-effects models using Eigen and S4. R package version 1.1-20.

[15] R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna. R Foundation for Statistical Computing.

[16] Drager, K., Hay, J. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change,* 24(1), 59-78.

* This study was supported by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209) and the Center for Asian and Pacific Studies at Seikei University.