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

XXX

XXX
XXX

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 and also 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 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 a shift of dominant cue from VOT to F0 is more advanced in accented words and is led by 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 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 the voice quality of the following vowel, as well as VOT. The pitch (F0) is higher when the vowel follows a word-initial voiceless than voiceless stops [4-7] and the voice quality of the vowel, as measured by h1-h2, is breathier (i.e., higher h1-h2) for the voiceless than the voiced condition [5, 6].

Studies also probed the perceptual cues listeners use to distinguish the initial stop voicing. [8] examined the perception of a monosyllabic word pair manipulated to vary in VOT (positive VOT values only) and F0 and found that the effect of F0 is visible only for very short VOT values. The participants in the study were all college students and came from four dialect regions, one of which was Kanto, which includes Tokyo. [9] created two stimuli sets representing different pitch accent conditions (*pasu* 'pass' vs. *basu* 'bus' for initial accented words, henceforth #HL; *teki* 'enemy' vs. *deki* 'result' for unaccented words, henceforth #LH). The stimuli were manipulated to vary in F0 and VOT, covering both the positive and the negative VOT values. They found that F0 plays a relatively minor role, but a more pronounced effect was found for the initial H than the initial L condition, which mirrors [7]'s finding that in production, the effect of initial stop voicing on F0 is stronger for H-initial than L-initial words. However, their stimuli were designed to cover a wider range for the initial H than the initial L series, and we cannot tell whether the interaction of pitch accent and F0 cue use is due to the larger difference pitch in the H-initial stimuli, rather than the listeners weighing the F0 cue more in the H-initial condition.

Finally, [1] is the only study that we know of that examined the age-based variation in perception. The majority of the participants in the study were from Tokyo but included speakers from other dialect regions. The study used naturally produced tokens of stops and 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. 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 older speakers were sparsely represented in the study overall, and we cannot rule out the possibility that the low count and the limited distribution of errors by older listeners were merely a function of the low number of participants.

Our study builds on these previous studies to examine what perceptual cues are used by Tokyo speakers to identify word-initial voiced and voiceless stops and how the cue use varies as a function of speakers' age and gender, as well as pitch accent.

Given the report that the change in Tokyo Japanese is led by younger females (cf. [10]) and the VOT cue is weakening in the speech of younger and/or female speakers, and the talker's own production pattern affects their own perception, we expect the sound change to be reflected in our data. In other words, other things being equal, we predict that female and younger speakers will show more innovative cue weighting and pay more attention to the vocalic cues than male and older speakers.

As for pitch accent, given that a larger F0 difference between voiced and voiceless stops is produced in the H-initial condition than in the L-initial condition, there are three possible ways F0 and pitch accent interacts in perception. One possibility is that if the range of F0 variation in the perceptual stimuli is kept comparable across the pitch accent conditions, listeners may give equal weights to both H and L pitch accent conditions. The second possibility is that given the relative salience of F0 cues in the H-initial condition, listeners may weigh F0 cues more for the H-initial words than for L-initial words. The third possibility is that listeners are less sensitive to F0 cues for the H-initial condition, requiring a larger F0 difference to shift the voicing boundary for H-initial words to match the large production difference in production.

2. METHODS

2.1. Perception experiment

Stimuli were two minimal pairs, 手前 [temae] vs.出前 [demae] (unaccented, #LH) and 天使 [tenɕi] vs. 電子 [denɕi] (initial accented, #HL). Coronal stop-initial words were chosen to avoid labials, which tend to include English loans, or dorsals, which tend to show less overlap in VOT between voiced and voiceless stops [2, 7]. Also, these words are commonly used and judged to be familiar to native speakers of (Tokyo) Japanese [11].

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

The duration and the F0 contour of the rest of the word were also manipulated to the average of all tokens for that pair to remove 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 average value for the speaker.

The manipulation parameters and the produced range for each acoustic dimension for each accent condition are summarized in Table 1, which were 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 the negative values and at 10 ms intervals for the 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 that are typical for each word were chosen. The three acoustic dimensions were orthogonally varied for each word pair to create 240 stimuli (=10 VOT steps * 6 F0 steps * 2 (h1-h2) baseline vowels * 2 pitch accent word pairs). Manipulations were done 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 (hailing from Tokyo, Chiba, Saitama, or Kanagawa) were recruited through an online crowdsourcing recruitment site (crowdworks.jp). A total of 172 speakers participated. Four speakers were excluded for answering "no" to the question "Do you speak Tokyo-style Japanese?" and 28 additional speakers who answered "yes" to this question but also listed other dialects they speak were excluded. Table 2 shows the breakdown of the 140 speakers included in the analysis by age and gender.

The task was word identification, whereby participants heard stimuli and chose the word they heard. The full experiment included informed consent, a background questionnaire, a production task, and a perception task, and it took 19.5 minutes on average. The perception experiment alone took 9.4 minutes.

2.2. Statistical analysis

For statistical analysis, we built a logistic mixed-

effects regression model using the lme4 package [13] in R [14] that 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. The phonetic variables, VOT (ms), F0 (st), and BASE.VOWEL (voiced = 0, voiceless =1) were z-score transformed to put them on a comparable scale. ACCENT was sum-coded (voiced = -0.5, voiceless = 0.5). Also included were a by-PARTICIPANT random intercept and by-PARTICIPANT random slope adjustments for the three phonetic predictors. The formula is shown (1).

(1) `RESPONSE ~ VOT x F0 x BASE.VOWEL x ACCENT x YOB x GENDER + (1 + VOT + F0 + BASE.VOWEL | PARTICIPANT)`

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 the voiceless BASE.VOWEL all lead to more voiceless responses. We can also see that these predictors interact, and the statistical model shows a significant four-way interaction (VOT x F0 x BASE.VOWEL x ACCENT). In other words, the predictors' effects are not uniform across different contexts.

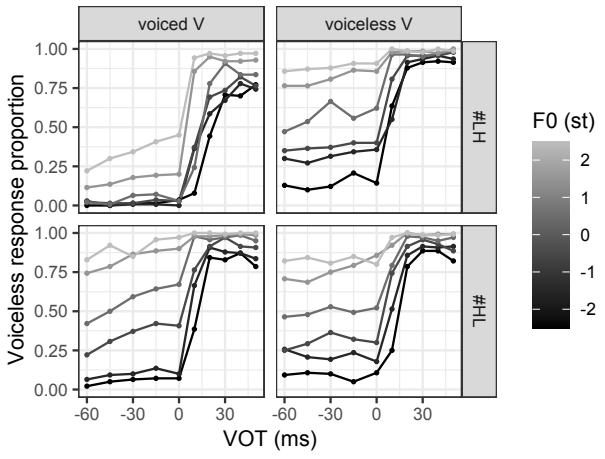

Figure 1: Proportion of voiceless responses by VOT, F0, BASE.VOWEL, and ACCENT, across all participants

Figure 2. breaks down the data by the participants' age and gender. While the overall patterns are similar regardless of age or gender, there's variation. In the statistical results, the six-way interaction of all predictors was significant.

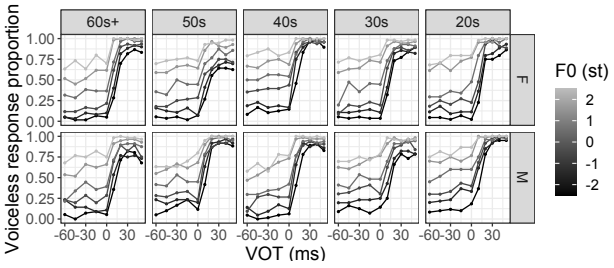

Figure 2: Proportion of voiceless responses by VOT, F0, AGE, and GENDER

	VOT	F0	vowel
main effects	2.098	1.591	0.647
2-way interaction			
x pair (#HL-#LH)	-0.573	0.326	-1.631
x gender (M-F)	(-0.054)	-0.239	0.137
x year of birth (yob)	(-0.007)	(0.053)	(0.012)
3-way interaction			
x pair x gender	(-0.165)	-0.241	(-0.011)
4-way interaction			
x pair x gender x 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. Parentheses indicate non-significant effects.

We will not interpret all interactions in this paper and instead focus on the main effects of the three phonetic predictors and their interaction with non-phonetic predictors only. These interactions tell us how the cue use changes depending on the pitch accent condition and the participant's age and gender. Table 3 summarizes the coefficient estimates. Interactions involving two or more non-phonetic predictors are shown only for those with significant effects. The scatterplots in Figure 3. show the distribution of individual participants' estimated slopes for the phonetic cues calculated from the regression model.

The main effects of all three phonetic predictors are found to be significant, but they interact with ACCENT significantly, which indicates that the 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 x -0.5) = **2.385**) than F0 (β = **1.428**) or 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 [tenɛi]-[denɛi], 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, individuals' slopes are tightly clustered around 2 for F0, while VOT coefficients are more spread out, with some even falling below 0. BASE.VOWEL coefficients are distributed around 0, which means that the cue does not have a consistent effect and that many participants use the cue in the opposite direction.

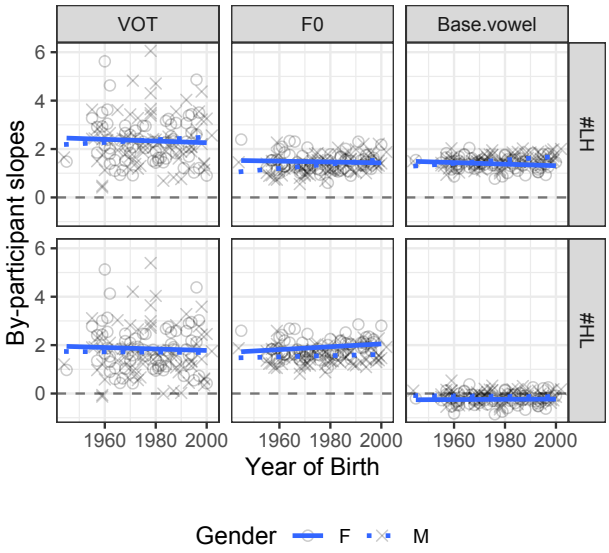

Figure 3: By-participant slope estimates for phonetic predictors, plotted by ACCENT and by the participants' GENDER and YOB. The dashed horizontal lines mark the slope of 0 and the solid and dotted lines are linear smooths for female and male participants.

As for the speaker-level predictors, age and gender, if individuals' perception reflects the sound change in progress whereby 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 x GENDER, F0 x ACCENT x GENDER, and F0 x ACCENT x GENDER x YOB shows that the F0 cue is stronger for female than male, and this interaction is stronger for the #HL [tenɕi]~[denɕi] pair and for the 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. 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 the word pair in the model. For the unaccented (#LH) pair, VOT is still the dominant cue regardless of the 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 chose 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 is shifting from VOT to F0 and the change is more advanced in the female participants and for the 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' best phonetic cue (=highest coefficient) by the ACCENT and the participant's age and gender. * indicates the majority pattern for the demographic subgroup.

4. DISCUSSION

In this paper, we aimed to examine the perceptual cue use for Tokyo Japanese stop voicing in the context of sound change in progress. We found that pitch accent had a significant effect and the cue weighting pattern reflects the production difference in the stimuli talker's speech (and likely that of Tokyo Japanese speakers' speech more generally). In the talker's production (Table 1), F0 differed by voicing more for the #HL pair than the #LH pair, while VOT and BASE.VOWEL quality differed more in the #LH than the #HL pair. Perception mirrors this pattern. We also found the effect of listeners' age and gender in the expected direction. For the #HL word pair only, female and younger listeners relied more heavily on F0 than male and older listeners. This interaction of F0 with age and gender affected the relative importance of VOT and VOT cues were relied on relatively less than F0 by younger female listeners. In short, the dominant cue for the initial voicing contrast is shifting from VOT to F0, and the shift is led by #HL words and by younger females.

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] Labov, W. 2001. *Principles of Linguistic Change, volume 2: Social Factors*. Blackwell Publishers.

[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] 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.

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