VOT-F0 COARTICULATION IN JAPANESE: PRODUCTION-BIASED OR MISPARSING?

Jiayin Gao[1,2,3], Jihyeon Yun[1,4], Takayuki Arai[1]

[1]Sophia Univ., [2]JSPS, [3]LPP, LaCiTO, LabEx EFL-Univ. Paris 3, [4]Chungnam National Univ.

ABSTRACT

In production, word-initial voicing contrast of plosives in Tokyo Japanese is not robustly based on VOT, since young speakers tend to devoice previously voiced plosives. Meanwhile, speakers rely heavily on f0. The present study aims to examine the role of VOT and f0 cues in perception. We conducted an identification test using resynthesized stimuli along a VOT continuum (-60 to +40 ms) orthogonal to an f0 continuum. Results suggested a categorical perception of VOT, while f0 was especially useful when VOT was around 0 ms. Higher f0 contours affected the response rate more than lower f0 contours, suggesting that the perceptual role of f0 raising was more important than f0 lowering.

Hence, plosive devoicing in production is not preceded by listeners' misparsing of low f0 and prevoicing. Their misattribution of high f0 to voicelessness may play a more important role in the shift of VOT-f0 cue weighting in perception and production.

Keywords: Japanese, voicing, perception, VOT, f0, coarticulation, tonogenesis

1. INTRODUCTION

1.1. Production-perception link in tonogenesis

Transphonologization [6] in the domain of tone has been widely documented. A well-known example is the loss of a consonantal voicing contrast replaced with a tonal contrast on the following vowel. This tonal development has been accomplished in many Southeast and East Asian languages [8]. One of the classic sound change models is *perception-biased* [14]: sound change is initiated by the listeners' *misparsing* of a sequence, attributing more weight to the coarticulatory effect than to the coarticulatory source in *perception*. According to this model, we may distinguish three stages in a tonal development: (1) presence of both the coarticulatory source and effect, that is, prevoicing of C, and lower f0 (fundamental frequency) on following V, presumably, (2) shift in *perception* from the coarticulatory source to effect, and (3) shift in *production* from the coarticulatory source to effect.

Such tonal development in progress, in an earlier or later stage, has been reported in other genetically and/or geographically unrelated languages, such as Afrikaans [4], Dutch [15], Tamang [12,13]. These studies also compared production with perception data and concluded that perception lags behind production, contrary to Ohala's model [14]. In particular, (some) speakers who use primarily f0 and devoice previously voiced plosives in production still rely heavily on prevoicing in categorization of these plosives. The most studied incipient tonogenetic case is Seoul Korean, in which two of the three plosive series are merging in terms of VOT (voice onset time) and the contrast is now realized on the f0 of the following vowel [19,9,1]. The tonal contrast is, however, not yet phonologized, as argued in these studies. In [1], it is also suggested that the tonogenesis is *production-biased*, driven by the VOT contrast reduction due to lenition in high-frequency words, combined with an adaptive expansion of f0 contrast.

There is no doubt that both production and perception play important roles in a sound change, and it is highly possible that one does not strictly follow the other. In this study, we attempt to gain some insight on the production-perception link from a situation prior to incipient tonogenesis.

1.2. This study

Tokyo Japanese possesses two plosive series traditionally described as prevoiced and voiceless unaspirated. However, recent studies suggested a trend towards devoicing word-initial voiced plosives by young speakers [20,5]. Meanwhile, VOT measures indicate that word-initial voiceless plosives are moderately aspirated [18,17,5]. Importantly, VOT of the two plosive series overlap [5]. As a secondary cue, f0 of the following vowel is higher after word-initial voiceless than voiced plosives, the magnitude and duration being more important in an H tone than L tone mora [5]. While it remains unclear whether aspiration of voiceless plosives or f0 difference on the following vowel is undergoing any recent change, devoicing of voiced plosives is argued to be an ongoing phonetic change. That is, the coarticulatory source is disappearing in Tokyo Japanese, similarly to Afrikaans, Dutch, and Tamang (see §1.1).

If this synchronic voicing variation may trigger a diachronic change involving exaggeration of the coarticulatory effect and complete loss of the coarticulatory source, can we already observe a shift from VOT to f0 cue in perception? To address this question, we conducted a forced-choice identification test in order to examine Japanese listeners' sensitivity to these two cues.

2. EXPERIMENT

2.1. Method

2.1.1. Participants

Nineteen native speakers (7 males, 12 females) of Tokyo Japanese living in Tokyo, with a mean age of 23 (from 20 to 31), participated in this experiment. None reported any speech or hearing disorder.

2.1.2. Resynthesized stimuli

Stimuli were created by modifying two minimal pairs produced in isolation by two males (aged 21 and 22): one minimal pair with HL tone (or pitch-accent), /pasu/ 'pass' vs. /basu/ 'bus' (loanwords, for lack of better choices), the other with an LH tone, /teki/ 'enemy' vs. /deki/ 'result, performance'. /b/ and /d/ were originally produced with prevoicing. For each word pair, two orthogonal continua were constructed varying VOT and f0.

Each VOT continuum was composed of 8 stimuli from -60 ms to +40 ms (step 1 to 8), with a step of 15 ms in the negative range [-60, -15] and of 10 ms in the positive range [10, 40]. /pasu/ and /teki/ were chosen as the base tokens, with original VOT at 35 and 30 ms for /p/ and /t/, respectively. For all stimuli, the original release part of /p-t/ was weakened so as to sound more natural when prevoicing was added. For positive VOT stimuli, the aspiration part of /p-t/ was lengthened or shortened by *Duration manipulation* in the PSOLA [21] program implemented in Praat [3]. For negative VOT stimuli, prevoiced portions extracted from original /basu/ and /deki/ were spliced before stimuli with a shortened VOT of 10 ms (approximately the release duration). The prevoiced portions, originally at 143 and 65 ms for /b/ and /d/, respectively, were then shortened to meet each of the 4 steps in the negative VOT range.

An equidistant f0 continuum composed of 6 stimuli from lowest to highest (steps 1 to 6) was imposed on the first vowel of each VOT-manipulated stimulus, using *Pitch manipulation* in the PSOLA [21] program implemented in Praat. (The second vowel kept the original f0 contour of the base tokens, unaffected by onset voicing.) The f0 contours of steps 2 and 5 were stylized from the natural f0 contours of /basu, deki/ and /pasu, teki/, respectively. Steps 3 and 4 were intermediate contours interpolated between steps 2 and 5 in equal f0 (Hz) space. Steps 1 and 6 were endpoint contours extrapolated from steps 2 to 5 in equal f0 (Hz) space. The f0 difference is larger in an initial H tone than L tone mora in natural productions, leading to a larger f0 range for resynthesized /p-ba/ (H tone /a/: 111–154 Hz) than for /t-de/ (L tone /e/: 114–133 Hz).

In total, 96 stimuli (8 VOT steps × 6 f0 steps × 2 tones) were constructed.

2.1.3. Procedure

The experiment was conducted using an adapted Python program [22]. Participants were tested individually in a soundproof room, following oral and written instructions in Japanese. Auditory stimuli were presented to them through a professional quality headphone and an Audio interface (Edirol) connected to a laptop computer. At trial onset, a fixation cross was displayed at the centre of the screen; 500 ms later, one auditory stimulus was presented; at stimulus offset, the fixation cross was replaced with two visual stimuli representing the two possible responses written in Japanese at the left and right side of the screen. The response side was the same for each listener for facilitation but counterbalanced across all listeners. Listeners were instructed to respond as quickly and accurately as possible by pressing the left or the right <SHIFT> key on the keyboard. The response time-out was set to 2 seconds. The test phase was preceded by a training phase of 10 trials consisting of original stimuli with unambiguous prevoicing or aspiration, and f0 pattern, during which listeners were given feedback for their correctness and response time, measured from the onset of visual stimuli presentation. During the test phase, the stimuli were presented in a different randomized order for each listener. For 14 listeners, each stimulus was repeated 3 times only due to the time schedule, yielding 288 trials in total,. For the other 5 listeners, each stimulus was repeated 4 times, yielding 384 trials in total. The test took about 15 to 20 minutes.

2.2. Results

2.2.1. Identification

Voiced identification curves are plotted as a function of VOT steps (Figure 1), and as a function of f0 steps (Figure 2). Figure 1 shows clear S shape curves in most of the cases, indicating a highly categorical perception of VOT. For /p-ba/ (H tone), the identification curves for the three highest f0 steps (4-6) show a shift towards the leftmost VOT endpoint as f0 increases, whereas the curves for f0 steps 1-3 are superimposed. In particular, f0 step 6 noticeably lowers voiced response rate at the leftmost VOT endpoint whereas the other steps do not. For /t-de/ (L tone), the identification curves for each f0 step are close to each other. There is a noticeable shift between f0 step 6 and the other steps. These observations suggest listeners' higher sensitivity to high pitch (steps 4-6, notably 6) than low pitch.

Figure 1: Voiced response rate as a function of VOT step, for the six f0 steps, for H tone (upper panel) and L tone (lower panel).

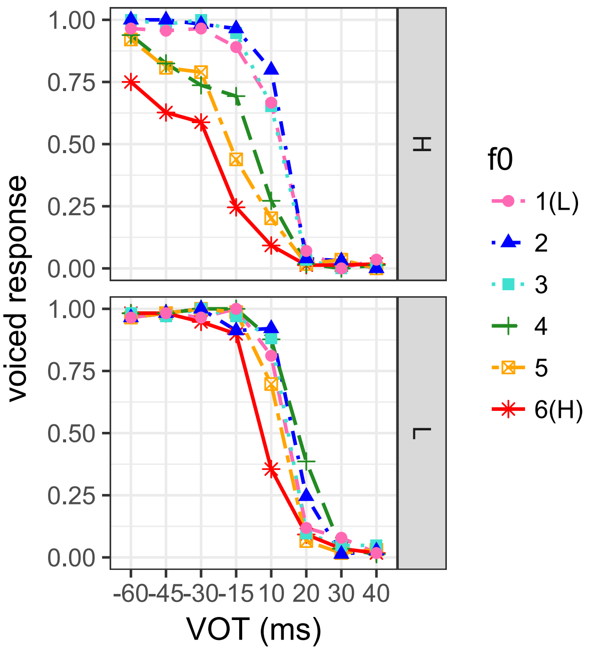
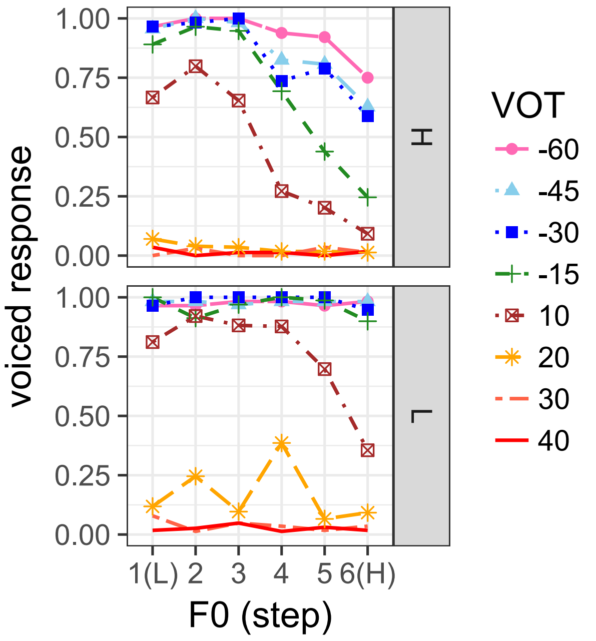


Figure 2: Voiced response rate as a function of f0 step, for the eight VOT steps, for H tone (upper panel) and L tone (lower panel).



In contrast, Figure 2 shows flat curves in most of the cases, indicating that the role of f0 is only secondary. It can be noticed that for L tone, f0 plays a role in voicing judgement only when VOT is at 10 ms. For H tone, the role of f0 can be clearly observed when VOT is at 10 and -15 ms. When VOT is below -15 ms, only the highest f0 steps affect voicing judgement.

Furthermore, mid-long lag VOT (≥20 ms for H tone, ≥30 ms for L tone) entails voiceless responses, regardless of f0. This suggests that aspiration is a reliable perceptual cue for voicelessness.

A generalized linear mixed model (GLMM) was fitted to the binomial voiced/1-voiced response data, using the *lme4* package [2] in R [16]. The following predictors were included in the model: *VOT (step, numerical)*, *f0 (step, numerical)*, *tone* (L as the reference level), and *f0×VOT×tone* interatction. The model also included random intercepts for *participant*, and *by-participant* random slopes for *f0*, *VOT*, and *tone*. Table 1 shows the results of the GLMM model. *f0×VOT* interaction has a significant effect only for H tone, but not for L tone.

Table 1: Results of GLMM fit to the response.

Predictor	Estimate	z value	Pr(>|z|)
(Intercept)	12.28	1.25	< 2e-16 ***
VOT	-2.11	-10.27	< 2e-16 ***
f0	-0.38	-1.70	0.090.
toneH	2.45	2.06	0.039 *
VOT:f0	0.03	0.69	0.488
VOT:toneH	-0.51	-2.32	0.020 *
f0:toneH	-1.47	-5.55	2.88e-08 ***
VOT:f0:toneH	0.21	4.22	2.45e-05 ***

Table 2 shows the 50% crossover boundaries (in step number) for voiced responses of the logistic curves calculated with the above GLMM model. The boundary is shifted towards the VOT rightmost endpoint as f0 decreases, for both H and L tones. The distance of the shift is greater for f0 steps 4-6 than 1-3 for H tone.

Table 2: 50% crossover boundaries (in VOT step number), for each f0 steps for H and L tone.

tone/f0step	6 (H)	5	4	3	2	1 (L)
H	3.0	3.8	4.4	4.8	5.1	5.4
L	5.1	5.3	5.4	5.5	5.6	5.7

Post-hoc pairwise comparisons were made for two separate GLMM model for H and L tone, with the same predictors as above but f0 step as a factor variable, using the *emmeans* package [11] in R, p-values adjusted with the Tukey method. Significant consecutive step contrasts at VOT step=4.5 are summarized in Table 3. It again shows an effect of f0 between higher f0 but not between lower f0 steps.

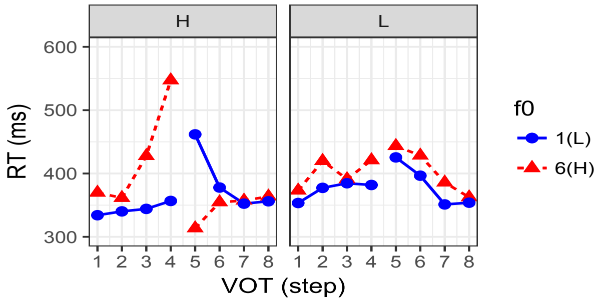
Table 3: Results of Post-hoc pairwise comparisons for f0 step contrasts.

tone	f0 step contrast	Estimate	z.ratio	p.value
H	step 3 – step 4	2.60	7.62	< .0001
	step 5 – step 6	0.84	3.46	0.007
L	step 4 – step 5	1.26	3.05	0.028
	step 5 – step 6	0.98	3.32	0.011

2.2.2. Response time

Response time (RT) data was analysed in order to examine whether conflicting cues (e.g., negative VOTs and high f0) slowed down listeners' responses. RT data was positively skewed. Hence, outliers (>Q3+1.5*IQR, 3.2% of the data) were removed, resulting in a relatively normal distribution. We used RT data for voiced responses within negative VOTs (steps 1-4), and for voiceless responses within positive VOTs (steps 5-8). For a clearer comparison of f0, only patterns for the lowest f0 (step 1) and the highest f0 (step 6) are shown in Figure 3.

Figure 3: RT (for dominant responses) as a function of VOT step, for the highest and lowest f0 steps, for H (left panel) and L tone (right panel).



For H tone, conflicting cues clearly lengthened RT for VOT steps 3-5, that is, the ambiguous VOT range for voicing judgement. Indeed, post-hoc pairwise comparisons showed an effect of f0 (steps 1 vs. 6) only for VOT at -30 ms [*t*=2.5, *p*=.01], -15 ms [*t*=4.0, *p*=.0001] and 10 ms [*t*=-3.4, *p*=.0008]. For L tone, RT had little effect.

3. DISCUSSION

This study examined the role of VOT and f0 in perception of word-initial plosive voicing in Tokyo Japanese. We questioned whether the ongoing devoicing of previously voiced plosives in production is led by a shift from VOT to f0 cue in perception.

Our identification and RT results showed that, in perception, listeners rely heavily on VOT and secondarily on f0. However, VOT strongly interacts with f0, as shown in the following observations: (1) mid-long lag VOT almost always entails voiceless responses, regardless of f0, suggesting that moderate aspiration is sufficiently reliable to cue voicelessness; (2) prevoicing entails a high percentage of voiced responses; (3) in an H tone mora, when prevoicing conflicts with high f0, listeners are biased towards voiceless judgement; and (4) when VOT is ambiguous (around 0 ms), listeners rely heavily on f0, being more sensitive to high f0 than low f0.

No evidence suggests a misparsing between prevoicing and low pitch, for listeners are highly sensitive to prevoicing, and do not attribute low pitch to voicedness when VOT is ambiguous. Instead, they tend to attribute high pitch to voicelessness. If this is a misparsing process which may eventually lead to a sound change, it might suggest that, the coarticulatory source is voicelessness and the coarticulatory effect is high f0. In fact, recent studies on production data of obstruent voicing in English, French and Italian proposed that f0 perturbation is due to an f0 raising effect of voiceless obstruents rather than an f0 lowering effect of voiced obstruents as commonly accepted [7,11]. Nonetheless, it is important to note that this misparsing occurs especially (but not only) when VOT is ambiguous.

Let's now come back to the question of devoicing in production. If prevoicing is helpful in perception, why, then, isn't it regularly produced? The only explanation we can think about is a context-induced production bias. Voiced plosives in domain initial position (word/phrase- and especially utterance initial) might be simply more prone to devoicing, as observed in English, as well as several languages reported to be undergoing tonal development (see §1). Note that in these languages, prevoicing is well preserved in word-medial position. It is also interesting to note that in a language like French, in which prevoicing is a robust cue, lexical words are rarely phrase-initial, often preceded by functional words.

In conclusion, in Tokyo Japanese, variable plosive voicing in word-initial position can be viewed as a synchronic variation similar to English, resulting from a context-induced production bias of devoicing. Subsequently, VOT becomes less reliable and f0 contrast is expanded to recover/disambiguate VOT contrast reduction, possibly due to an f0 raising rather than f0 lowering process, in perception as well as in production.

Whether this synchronic variation is a precursor of a possible tonal development depends on multiple factors beyond the scope of this paper.

4. ACKNOWLEDGEMENT

This study is funded by JSPS KAKENHI Grant Number 17F17006.

5. REFERENCES

[1] Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M., Yoon, T.-J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120-144.

[2] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2017). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-14.

[3] Boersma, P., & Weenink, D. 2017. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.33.

[4] Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., Wissing, D. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics*, *66*, 185-216.

[5] Gao, J., Arai, T. 2018. F0 perturbation in a "pitch-accent" language. *Procs. 6th Int'l Symposium on TAL Berlin,* paper n. 13, 1-5.

[6] Hagège, C., Haudricourt, A.-G. 1978. *La Phonologie Panchronique*. Paris: Presses Universitaires de France.

[7] Hanson, H. M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am.*, *125*(1), 425-441.

[8] Haudricourt, A.-G. 1961. Bipartition et tripartition des systèmes de tons dans quelques langues d'Extrême-Orient. *Bulletin de la Société de Linguistique de Paris*, *56*(1), 163-180.

[9] Kirby, J. P. 2013. The role of probabilistic enhancement in phonologization. In: Yu A. C. L. (ed.), *Origin of sound change: Approaches to phonologization.* Oxford: Oxford University Press. 228-246.

[10] Kirby, J. P., & Ladd, D. R. 2016. Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *J. Acoust. Soc. Am.*, *140*(4), 2400-2411.

[11] Lenth, R. 2018. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0.

[12] Mazaudon, M., Gao, J. 2018. Cue weighting after a tone-split in Tamang: A perception study of stop initial words. LabPhon16 Lisbon.

[13] Mazaudon, M., Michaud, A. 2008. Tonal contrasts and initial consonants: a case study of Tamang, a 'missing link' in tonogenesis. *Phonetica, 65*, 231-256.

[14] Ohala, J. J. 1981. The listener as a source of sound change. In: Masek, C. S., Hendrick, R. A., Miller, M. F. (eds.), *Chicago Linguistic Society: Papers from the Parasession on language and behavior*. Chicago: Chicago Linguistic Society. 178-203.

[15] Pinget, A.-F. 2015. *The actuation of sound change* Unpublished Ph.D. Dissertation. Utrecht University.

[16] R Core Team. 2017. R: A Language and Environment for Statistical Computing. R version 3.4.2. R Foundation for Statistical Computing.

[17] Riney, T. J., Takagi, N., Ota, K., Uchida, Y. 2007. The intermediate degree of VOT in Japanese initial voiceless stops. *Journal of Phonetics*, *35*(3), 439-443.

[18] Shimizu, K. 1996. *A cross-language study of voicing contrasts of stop consonants in six Asian languages.* Tokyo: Seibido.

[19] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, 23, 287-308.

[20] Takada, M., Kong, E. J., Yoneyama,K., Beckman, M. E. 2015. Loss of prevoicing in modern Japanese /g, d, b/. *Procs. 18th ICPhS Glasgow*, paper n. 873, 1-5.

[21] Valbret, H., Moulines, E., Tubach, J. P. 1992. Voice transformation using PSOLA technique. *Speech Communication* 11(2), 175-187.

[22] Yu, S. (2018). Psypsy. Github repository <https://github.com/shi4yu2/psypsy>