Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery

Jiayin Gao, Takayuki Arai

Plosive (de-)voicing and *f*0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery

Jiayin Gao,[abcd]* Takayuki Arai[a]

*Corresponding author: jiayin.gao@gmail.com
[a] Sophia University, 7-1 Kioi-chô, Chiyoda-ku, 102-8554 Tokyo, Japan
[b] Japan Society for the Promotion of Science, 5-3-1 Kôjimachi, Chiyoda-ku, 102-0083 Tokyo, Japan
[c] Laboratoire de Phonétique et Phonologie (CNRS – Univ. Sorbonne Paris Cité – LabEx EFL), 19 rue des Bernardins, 75005 Paris, France
[d] Laboratoire Langues et Civilisations à Tradition Orale (CNRS – Univ. Sorbonne Paris Cité – LabEx EFL), 7 rue Guy Môquet, 94801 Villejuif, France

Abstract
This study addresses the two-way laryngeal contrast of plosives in Tokyo Japanese, which is commonly analyzed as a "true voicing" language. We examine how voicing-related properties of the plosive and *f*0 of the following vowel varied with the position in the word and in the sentence. We compare word-initial with word-medial positions for words in citation (between two pauses) and for two prosodic conditions in a carrier sentence: with vs. without a preceding pause. In word-initial position, unlike in a typical "true-voicing" language such as French, voiced plosives in Tokyo Japanese show a high devoicing rate, while voiceless plosives are moderately aspirated. A combination of VOT and *f*0 of the following vowel is used to distinguish the two plosive series. In word-medial position, voiced plosives are frequently prevoiced and voiceless plosives are unaspirated, while *f*0 does not differ after the two plosive series. This positional variation suggests that the onset-induced *f*0 effect is enhanced in word-initial position, where the VOT cue is not sufficient, but not in word-medial position, where the plosive voicing contrast is robustly marked by presence vs. absence of phonetic voicing. The differential use of cues in different environments in Tokyo Japanese provides another piece of evidence for the complexity of phonetic implementations of the voicing contrast. Finally, we discuss the enhancement of *f*0 perturbations as a source of a potential tonal development and ask whether such a development would take place in Tokyo Japanese.

Keywords
Japanese, voicing, aspiration, VOT, *f*0 perturbations, cue weighting, cue enhancement

Plosive (de-)voicing and *f*0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery

1. Introduction

This is a phonetic study of voicing-related and *f*0 (fundamental frequency) properties associated with the laryngeal contrast of plosives in Tokyo Japanese. (Throughout the paper, we will also employ the term "cue" in a similar way as "property," without any perceptual implication.) Japanese is commonly described as having a "true voicing" contrast, as in French, rather than an "aspiration" contrast, as in English (e.g., Nasukawa, 2005). However, recent studies show a trend towards devoicing word-initial voiced plosives in several Japanese dialects including Tokyo Japanese (Takada, 2011; Takada, Kong, Yoneyama, & Beckman, 2015). How, then, is the voicing contrast realized in modern Tokyo Japanese?

In this introductory section, we shall review the phonetic implementations of laryngeal features proposed for two types of languages, "aspirating" and "true voicing" languages, with respect to voicing, aspiration, and *f*0 on the following vowel (Section 1.1). We shall then confront the traditional account of Tokyo Japanese as a "true voicing" language, where the [voice] feature is active, with recent phonetic data revealing an atypical "true voicing" pattern (Section 1.2). Finally, we shall present our research questions and hypotheses concerning the phonetic realizations of the voicing contrast in Tokyo Japanese, aiming to achieve a better understanding of the typology of the voicing contrast and its relationship with *f*0 of the following vowel (Section 1.3).

1.1 Voicing contrast and its relationship with f0

1.1.1 Phonetic and phonological specifications of the laryngeal contrast

The [+/-voice] feature is often employed to represent the most commonly found two-way laryngeal contrast in obstruents. Keating (1984) claimed this feature to be universal, and interpreted crosslinguistically diverse phonetic patterns as different phonetic implementations, or specifications, of this unique feature. In many Germanic languages, [+voice] is specified as {voiced} or {voiceless unaspirated}, and [-voice] as {voiceless aspirated}. In many Romance and Slavic languages, [+voice] is specified as {voiced}, and [-voice] as {voiceless unaspirated}. (Accolades { } are used to note phonetic features.) These two phonetic specifications are often referred to as "aspirating" (for the Germanic type) and "true voicing" (for the Romance and Slavic type) (e.g., Beckman, Jessen, & Ringen, 2013). The correspondence with language families is only indicative, because language and dialectal variations have been reported (e.g., van Alphen & Smits, 2004, for Dutch; Pape & Jesus, 2011, for European Portuguese; Caramazza & Yeni-Komshian, 1974, Sundara, 2005, for Canadian French).

In defining the phonetic specifications of plosives, V(oice) O(nset) T(ime) of the surface segments has been widely used as the main phonetic criterion (Lisker & Abramson, 1964). Plosives with a negative VOT (or voice lead) are specified as {voiced}. Plosives with a positive and long-lag VOT are specified as {voiceless aspirated}. Plosives with a positive and short-lag VOT are specified as {voiceless unaspirated}, or are unspecified in some "privative" models. It should be remembered that VOT is a measure of the timing relationship, which only *indirectly* indicates the abduction/adduction gesture during closure and the width of the glottal opening. Although VOT alone cannot, and was never intended to, capture all the aspects of plosive contrasts, these measurements are certainly much easier to carry out than articulatory ones (cf. Abramson & Whalen, 2017; Cho, Whalen, & Docherty, 2019) and will also be used in the present study.

The use of VOT in defining the phonetic specifications of plosives has several complications (see Abramson & Whalen, 2017, p. 81). One of them is that a language could choose an intermediate category between "true voicing" and "aspirating," as is evident in modern Hebrew, in which voiced plosives consistently have negative VOTs but voiceless plosives have medium-lag VOTs (Raphael, Tobin, Faber, Most, Kollia, & Milstein, 1995). Other complications include the fact that VOT may be affected by place of articulation, the following vowel, or prosodic strengthening. Prosodic strengthening may enhance syntagmatic contrast and/or paradigmatic contrast (for a review, see Cho, 2016, and references therein). For example, VOTs of domain-

initial {voiceless aspirated} plosives are longer in phrase-initial than in phrase-medial position, and longer in phrase-medial than in word-medial position in Korean (Cho & Keating, 2001). Similar results have been found in English (Pierrehumbert & Talkin, 1992), but they seem to be limited to unaccented positions (Kim, Kim, & Cho, 2018). In these cases, VOT lengthening can be interpreted as a manifestation of syntagmatic contrast enhancement by increasing the voicelessness of these plosives, and thus their consonantality. This may also be viewed as a manifestation of paradigmatic contrast enhancement by maximizing the phonetic distinction between plosives specified as {voiceless aspirated} and other plosive series. In contrast, in French, VOT of voiceless unaspirated plosives shows little variation between different prosodic positions (Fougeron, 2001). In Dutch, VOTs of these plosives are shortened in prosodically stronger positions, which is interpreted as an enhancement of the {voiceless unaspirated} feature (Cho & McQueen, 2005). On the other hand, VOTs of voiced plosives have received less attention. In Taiwanese, a longer voice lead, that is, a more negative VOT of {voiced} plosives in domain-initial than in domain-medial positions, contributes to a paradigmatic contrast enhancement (Hsu & Jun, 1998). In contrast, in American English, the voice lead of {voiceless unaspirated} plosives is found to be reduced or lost in prosodically stronger (domain-initial, accented, stressed) positions, which enhances a syntagmatic contrast, but enhances a paradigmatic contrast less systematically (Kim et al., 2018).

Previous crosslinguistic data have also shown that voiced plosives in domain-initial position have different VOT realizations depending on their phonetic specifications. Plosives specified as {voiced} are robustly prevoiced (i.e., with negative VOTs) in all positions, while those specified as {voiceless unaspirated} are frequently devoiced (i.e., with positive VOTs) in domain-initial position. For example, the percentage of prevoiced tokens in absolute initial position is 97% and 85.6%, in French and Spanish, respectively (Solé, 2018), remarkably higher than in English (Lisker & Abramson, 1967; Dmitrieva, Llanos, Shultz, & Francis, 2015) and German (Jessen, 1998), in which a great majority of them are devoiced. (See also Cho et al., 2019: Figure 2, for their review of the VOT data of 17 "true voicing" languages.)

Aside from prosodic influence, the presence or absence of a pause may also influence phonetic voicing depending on whether plosives are specified as {voiced} or {voiceless unaspirated}. Glottal vibration is difficult to initiate in post-pausal position but easy to maintain in intersonorant position. Hence, in some, but not all "aspirating" languages, {voiceless unaspirated} plosives undergo voicing in intersonorant position. This is argued to be a form of *passive voicing*, contrary to the *active voicing* of voiced plosives found in "true voicing" languages (Beckman et al., 2013). In this position, VOT measurements are theoretically irrelevant for a plosive preceded by a sonorant segment because voicing has started well ahead of the closure before the plosive. Instead of VOT, other acoustic measurements for phonetic voicing can be used, such as *percentage of voicing during closure* (Docherty, 1992), or *V(oicing)-ratio* (Snoeren, Hallé, & Segui, 2006), and *connection voicing* combined with *after closure time* (Mikuteit & Reetz, 2007). Beckman et al. (2013) compared the voicing pattern in intersonorant word-initial position without a preceding pause in German with previous reports on the pattern in Russian (Kulikov, 2012; Ringen & Kulikov, 2012), a "true voicing" language. The percentage of tokens with full voicing during the plosive in intersonorant word-initial position, which was defined as having over 90% voicing during closure, reached 97.5% in Russian, but it was only 62.5% in German. Since the voicing during closure in German was partial, the authors concluded that this voicing pattern is passive, lending support to an "aspirating" instead of a "true voicing" account.

Contrary to Keating's (1984) proposal, many other researchers, including Beckman et al. (2013), have treated the distinction between these two types of voicing contrast as a phonological property instead of a phonetic implementation. For Iverson and Salmons (1995), "This familiar typological difference between the majority of Germanic (weakly or passively voiced 'voiced' stops, aspirated voiceless stops) and the Romance and Slavic languages (thoroughly voiced voiced stops, unaspirated voiceless stops) is […] fundamental, a part of the phonological representation itself." They proposed the feature [spread glottis] to define the laryngeal contrast in "aspirating" languages, while reserving [voice] for this contrast in "true voicing" languages only. The authors further argued that [+spread glottis] (but not [-spread

glottis]) and [+voice] (but not [-voice]) are active in phonological processes — such as assimilation — in "aspirating" and "true voicing" languages, respectively. For this reason, a "privative" model is privileged, specifying the marked feature only. In their model, [spread glottis] contrasts with [ ] (unmarked) in "aspirating" languages, and [voice] contrasts with [ ] (unmarked) in "true voicing" languages.

1.1.2 Onset-induced f0 perturbations

Aside from VOT, multiple cues correlated to voicing have been reported, such as F1 transition, closure duration, energy of burst, and duration of the preceding vowel (Lisker, 1957; Peterson & Lehiste, 1960; Summerfield & Haggard, 1977; Repp, 1982; Serniclaes, 1987, among others). In particular, *f*0 of the vowel onset is higher when following a voiceless than a voiced obstruent. With English speakers/listeners, this onset-induced *f*0 perturbation effect is found in production data (House & Fairbanks, 1953; Lehiste & Peterson, 1961; Ohde, 1984) and plays a secondary role in perception (Haggard, Ambler, & Callow, 1970; Whalen, Abramson, Lisker, & Mody, 1993). This *f*0 perturbation effect has been attested crosslinguistically, in both "aspirating" and "true voicing" languages (Hombert, Ohala, & Ewan, 1979; for a review, see Hanson, 2009; Kirby, 2018; Cho et al., 2019, and references therein).

The magnitude of *f*0 perturbations in the languages reviewed by Coetzee and colleagues, including English, German, and French, ranges from 8 to 16 Hz (Coetzee, Beddor, Shedden, & Styler, 2018: Table 1). In languages analyzed as (possibly) undergoing incipient tonogenesis, the quasi-phonologized *f*0 difference is much greater. It may reach nearly 100 Hz for female speakers and nearly 50 Hz for male speakers of Seoul Korean (Silva, 2006: Figure 4), and about 40 Hz for young female speakers of Afrikaans (Coetzee et al., 2018: Figure 3). Regarding the duration of *f*0 perturbations, they generally disappear no later than the vowel midpoint. In some tone languages, it has been found that *f*0 perturbations are more limited in duration, for example, immediately after the stop release in Thai (Gandour, 1974), only at the vowel onset in Yoruba (Hombert et al., 1979). *f*0 perturbations also interfere with intonation, in that the magnitude/duration is generally larger in high than low *f*0 contexts (Kohler, 1982; Hanson, 2009; Kirby & Ladd, 2016). For tone languages, contradictory results have been reported concerning the interaction between the size of the *f*0 effect and the tone level. Kirby (2018) found larger *f*0 perturbations in the high-falling tone than in other tone contexts in Thai, whereas an opposite effect was found in Beijing Mandarin (Xu & Xu, 2003), and in word-medial position in Shanghai Chinese (Chen, 2011). Other prosodic effects have been reported. For example, focalization is generally associated with a larger *f*0 perturbation effect (Chen, 2011, for Shanghai Chinese; Hanson, 2009, for English; Kirby & Ladd, 2016, for French and Italian). Similarly, in isolated context, in part due to hyperarticulation, *f*0 perturbations are larger than in carrier sentence context. Indeed, Kirby (2018) reported a solid *f*0 perturbation effect in isolated context and an attenuated or absent effect in carrier sentences in Thai, Khmer, and Vietnamese.

The onset-induced *f*0 perturbation effect is crucial in the understanding of some tonal developments, such as tone split in many East Asian and Southeast Asian languages (Haudricourt, 1961), at the end of which a tonal contrast replaces a previous segmental contrast, such as a voicing contrast of the syllable onset. Hence, in the course of tone split, the *f*0 difference after voiceless vs. voiced onsets is necessarily enhanced so that it can be perceived and phonologized. A more fundamental question is the origin of this *f*0 perturbation effect prior to the tonal development.

Aerodynamic and articulatory properties have been proposed as the source of *f*0 perturbations. While they can be superimposed (Kohler, 1985), the aerodynamic account alone does not explain *f*0 perturbations that extend up to 100 ms (Hombert et al., 1979). At the articulatory level, *f*0 lowering after voiced compared to voiceless obstruents may be attributed to the lowering of the larynx during closure (Ewan, 1976). Larynx lowering, which helps to facilitate voicing during closure, may lead to a rotation of the cricoid cartilage, at least in a speaker's low *f*0 ranges, and consequently result in *f*0 lowering (Honda, Hirai, Masaki, & Shimada, 1999). On the other hand, during the closure of voiceless obstruents, tension in the cricothyroid musculature (CT) contributes to stiffening the vocal folds, thus inhibiting phonation during

voiceless obstruents (Halle & Stevens, 1971; Löfqvist, Baer, McGarr, & Story, 1989). Consequently, *f*0 is raised at the onset of the following vowel. This is argued to be the source of *f*0 perturbations in English (Hanson, 2009). In her acoustic study, Hanson (2009) used sonorant onsets as a baseline and found that *f*0 was raised after voiceless obstruents as compared to voiced obstruents and, more importantly, to the sonorant reference. Kirby and Ladd (2016) found similar results in French and Italian and concluded that *f*0 raising due to voicelessness is the main source of *f*0 perturbations in both "aspirating" and "true voicing" languages. On the other hand, another theory holds that *f*0 perturbations are *controlled*: that *f*0 after voiced obstruents is lowered *intentionally* by speakers for the purpose of enhancing the auditory percept of [voice] by reinforcing the low-frequency energy of the vowel onset (Kingston & Diehl, 1994).

Recent theories have come to a hybrid model reconciling the automatic and the controlled theories (Hoole & Honda, 2011; Dmitrieva et al., 2015), according to which *f*0 perturbations are automatic and biomechanical by nature, due to the aerodynamic and articulatory properties explained above, but *can be* enhanced deliberately to increase the distinctiveness of the phonological contrast. Dmitrieva et al. (2015) reported a negative correlation between VOT and onset *f*0 in voiceless plosives in English, suggesting that speakers enhance one cue to compensate for the weakening of the other cue. Such a negative correlation did not occur in either plosive series in Spanish, or in the prevoiced or devoiced plosives in English (also see Shultz, Francis, & Llanos, 2012; and Kirby & Ladd, 2015, for similar results). They argued that negative VOT could be perceptually salient and thus did not need *f*0 enhancement, but that long-lag VOT could be perceptually confused with short-lag VOT, thus motivating the deliberate use of *f*0 as an enhancing cue.

To sum up, VOT-*f*0 covariation is widely attested as a synchronic pattern, and it has also resulted in diachronic tonal developments in many languages. The source of this covariation has been attributed to automatic mechanisms — articulatory, and maybe aerodynamic — or to controlled enhancements, or to a hybrid effect of both. It should be noted, however, that there are counterexamples to this onset-induced *f*0 perturbation effect. Gordon (2016) found that *f*0 perturbations after voiced vs. voiceless obstruents were very limited and unsystematic in several American indigenous languages. In some languages, aspirated plosives are followed by a lower *f*0 than unaspirated plosives (see Chen, 2011, for a review). Ladd and Schmid (2018) further pointed out that *f*0 perturbations are not solely dependent on voicing or aspiration as broadly defined by VOT, but possibly on the articulatory strategies for voicing and for aspiration.

1.2 Voicing and f0 in Japanese

1.2.1 Voicing and f0 perturbations

Japanese has long been analyzed as having a voicing contrast. Itô and Mester (1986), followed by others, argued that [voice] was an active and privative feature, because it participates actively in a certain number of (morpho-)phonological rules in Japanese (Itô & Mester, 1986; Nasukawa, 2005). One of these rules is *Lyman's Law* (Lyman, 1894) in the *Rendaku* process. *Rendaku* is a morpho-phonological process for compounding in which the initial obstruent of the second element of the compound word, if voiceless, becomes voiced. *Lyman's Law* is the blocking of the *Rendaku* process when the second element of the compound word contains a segment *specified* as [voice]. (See Vance, 1987, ch. 10, for other rules involved in *Rendaku*.) As illustrated in (1), *Rendaku* applies in (1a) and (1b) but is blocked in (1c) because *kotoba* contains /b/, which is specified as [voice]. A sonorant /m/ or /r/ does not block *Rendaku*, because it is *underspecified* for [voice]. (Examples 1b and 1c are taken from Itô & Mester, 1986.)

(1a) ori 'fold' + kami 'paper' -> ori**g**ami 'paper folding'
(1b) onna 'woman' + kokoro 'heart' -> onna**g**okoro 'feminine feelings'
(1c) onna 'woman' + koto**b**a 'word' -> onna**k**oto**b**a 'feminine speech'

Once the [voice] feature is analyzed as active in these (morpho-)phonological rules, some phonologists presume that its phonetic implementation will be that of a "true voicing" language (Iverson & Salmons, 1995; Nasukawa, 2005). However, previous phonetic studies on Japanese do not show a typical "true voicing" pattern. Voiceless plosives in word-initial position are reported to have intermediate-lag VOTs. Shimizu (1996) described these plosives as "moderately aspirated," although the speakers he tested were Japanese-English bilinguals living in Edinburgh. Riney et al. (2007) measured VOTs on Japanese monolinguals and found similar results: 30.0, 28.5, and 56.7 ms for /p/, /t/, and /k/, respectively, in monomoraic words in a carrier sentence (Riney, Takagi, Ota, & Uchida, 2007).
As for voiced plosives, Takada and colleagues examined the evolution of the production of voiced plosives, using apparent-time and cross-dialectal data (Takada, 2011; Takada et al., 2015). Having compared young speakers with elderly ones, the authors concluded that in several Japanese dialects, including Tokyo dialect, in word-initial position in isolated words, young speakers tend to devoice these voiced plosives or produce a shorter voice lead than elderly speakers.
As for the relationship between voicing and *f*0, it has seldom been examined in Japanese. As in many other languages, *f*0 starts higher after voiceless than after voiced onsets. Shimizu (1996) reported a difference of about 29–38 Hz for females, and of about 7–18 Hz for males at the vowel onset after word-initial plosives in a carrier sentence. Kawahara (2006) reported a difference of about 20 Hz for three female speakers of Japanese living in the US, using nonce words in a carrier sentence.

1.2.2 Pitch-accent, or tone
The status of Japanese as a tone language is not uncontroversial. Japanese is traditionally labeled as a "pitch-accent" language. However, it may be viewed as a tone language, given that *f*0 is necessary for determining the meaning of a word in a number of tonal minimal pairs (Hyman, 2009), although this concerns a weak percentage of them (Shibata & Shibata, 1990). Here, we describe briefly the tone system of Tokyo Japanese. The tone bearing unit is the mora, but a tone pattern applies to a word or a phrase. For the sake of simplicity, we use the traditional categories under the "pitch-accent" label here. This allows us, quite simply, to describe a word or phrase as (a) accentless: *f*0 is gradually raised over the entire sequence, that is, the initial mora carries a lower *f*0 than the following moras; (b) carrying a non-initial accent: *f*0 is raised until a peak on the accented mora, drops abruptly on the following mora and remains low; or (c) carrying an initial accent: *f*0 starts high on the first mora, which is accented, drops abruptly on the following mora and remains low. For example, /*ha.si.ga*/ (*ga* is a subject particle) means 'edge' if it carries the LHH melody (accentless), 'bridge' if it carries the LHL melody (accent on the second mora), or 'chopsticks' if it carries the HLL melody (accent on the first mora).

1.3 Research questions and hypotheses
Tokyo Japanese presents an interesting example for the understanding of the following two aspects: (a) the phonetic implementation of the [voice] feature. The morpho-phonological rules in Tokyo Japanese support an active [voice] feature, yet the phonetic evidence is not clearly in favor of a "true voicing" pattern; (b) the *f*0 perturbation pattern. How is this related to the phonological voicing distinction and its phonetic specification?

To address these issues, the first goal of this study is to revisit the phonetic realizations of the two plosive series in Tokyo Japanese. These two plosive series are termed "voiced" and "voiceless" throughout the paper. Previous phonetic studies are limited to plosives in word-initial position, either in citation only or in a carrier sentence only. Our study will compare the production of the two plosive series between word-initial and word-medial positions, both in citation forms and in carrier sentence contexts. We aim to provide not only a comprehensive description of the phonetic realization of the two plosive series but also useful information about their phonetic specifications according to positions and prosodic conditions (see Section

1.1.1). We shall examine how the laryngeal contrast is implemented in Tokyo Japanese. We consider French a typical "true voicing" language, and English a typical "aspirated" language. A language could follow one or the other pattern, but it could also fall into an intermediate category, as does Hebrew (Raphael et al., 1995, see Section 1.1.1).

If Tokyo Japanese is a typical "true voicing" language, the following patterns can be expected:
(A1) In post-pausal position, voiced plosives are robustly produced with prevoicing, that is, voicing during closure, as indicated by negative VOTs;
(A2) In intersonorant position, voiced plosives are robustly produced with prevoicing, that is, a high percentage of voiced plosives are fully voiced;
(A3) Voiceless plosives are most likely unaspirated, as indicated by short-lag VOTs (mostly between 0 and 30 ms).

On the contrary, if Tokyo Japanese is a typical "aspirating" language, the following patterns can be expected:
(B1) In post-pausal position, voiced plosives are not robustly produced with prevoicing, but are most likely to be voiceless unaspirated, as indicated by short-lag VOTs (mostly between 0 and 30 ms);
(B2) In intersonorant word-initial position, voicing is, at most, passively maintained during the closure of voiced plosives, that is, a low percentage of voiced plosives are fully voiced;
(B3) Voiceless plosives are most likely aspirated, as indicated by long-lag VOTs (mostly longer than 50 ms);
(B4) VOTs of voiceless plosives are lengthened in domain-initial position.

If Tokyo Japanese lies somewhere between the two categories, we can expect it to follow some of the patterns A1-3 and some of the patterns B1-4.

The second goal of this study is to examine the *f*0 perturbation effect related to plosive voicing in Tokyo Japanese in different positions and prosodic contexts. In previous studies (Shimizu, 1996; Kawahara, 2006), *f*0 was measured at only one or two time points. In our study, the magnitude and the time course of *f*0 perturbations will be shown, and will be viewed in relation to the phonetic and phonological voicing of the plosive onset. Following Hanson (2009), and Kirby and Ladd (2016), *f*0 will be compared after plosive and /m/ onsets. *f*0 contour after sonorants is taken by these authors as a neutral baseline, because (a) sonorants do not participate in the phonological voicing contrast, and (b) *f*0 after sonorants is presumably not perturbed because of any automatic mechanisms. Hence, the use of a /m/ baseline will allow us to assess whether *f*0 perturbation is due to *f*0 raising after voiceless plosives or *f*0 lowering after voiced plosives.
We hope to get a better understanding of the source of *f*0 perturbations: Is this an automatic biomechanical effect or an enhanced cue to the voicing contrast in Tokyo Japanese? We are, of course, aware that it might be the case that these two sources are combined.

As reviewed in Section 1.1.2, there are at least two articulatory sources of *f*0 perturbations.
If the *f*0 perturbation effect is an automatic effect of *f*0 lowering due to the closure voicing of voiced plosives, the following patterns can be expected:
(X1) *f*0 is lowered after prevoiced plosives compared to the /m/ onset;
(X2) When voiced plosives are produced without closure voicing, *f*0 is raised; in other words, devoiced plosives are followed by higher *f*0 than prevoiced plosives;
(X3) *f*0 perturbations are observed in all positions and contexts;
(X4) Onset *f*0 correlates positively with VOT in the negative VOT range: the longer the prevoicing (i.e., smaller VOT), the lower the onset *f*0 value.

If the *f*0 perturbation effect is an automatic effect of *f*0 raising caused by voiceless plosives, the following patterns can be expected:

(Y1) *f*0 is raised after voiceless plosives compared to the /m/ onset;
(Y2) *f*0 perturbations are observed in all positions and contexts.

A controlled account of *f*0 perturbations implies that *f*0 perturbations are enhanced to contribute to the distinctiveness between voiced and voiceless plosives. Our view is that the distinctiveness is not necessarily achieved through a reinforcement of the auditory percept of "voicedness," as suggested by Kingston and Diehl (1994) (see Sections 4.2, 4.3, and 4.4 for discussion). If the *f*0 perturbation effect is due to a controlled enhancement, the following patterns can be expected:
(Z1) *f*0 perturbations are conditioned by the phonological voicing contrast, but not predicted by VOT: a similar *f*0 perturbation effect is observed regardless of the closure voicing of voiced plosives;
(Z2) *f*0 perturbations are larger in contexts in which the primary voicing cue is less reliable;
(Z3) Within each phonological voicing category, onset *f*0 correlates negatively with VOT: the smaller the VOT value, the higher the onset *f*0 value.

2. Data collection and methods

2.1 Participants

Eighteen native speakers (9 males and 9 females) of Tokyo Japanese participated in the recording. They were all born and raised in the broader Tokyo area, and have never spent more than two years in other regions or countries, except for one participant who lived in Hiroshima before entering elementary school in Tokyo. At the time of the recording (Oct. 2017-Feb. 2018), the participants were aged from 19 to 27 (with a mean age of 22.4). They were recruited from Sophia University (in Tokyo) and received a prepaid gift card for their participation. On the basis of their answer to two open-ended questions about their foreign language skills, two participants judged their English level as good, while all the others reported a basic or intermediate English level. None of the participants were advanced in any other foreign language. None reported any speech or hearing disorders. The present research project (No. 2017-63) was approved by the ethics committee of Sophia University. All participants signed a written informed consent form.

2.2 Speech materials and recording

The speech materials consisted of 37 (near-)minimal pairs of lexical words with a voicing contrast for the plosive onsets /p-b, t-d, k-g/, plus 4 words with /m/ onset as a reference for analyses, making a total of 78 words.[1] Within each word pair, words of high and similar degrees of familiarity were used, based on the NTT database (Amano & Kondo, 2000). Each word contained 2 to 3 syllables and 2 to 5 moras. The target C was either in word-initial position, that is, the onset of the first syllable, or in word-medial position, that is, the onset of the second syllable. The wordlist contained 38 accentless words and 40 initial-accent words (see Section 1.2.2 for a description of pitch-accent in Tokyo Japanese). In an accentless word, the initial mora had an L tone, and a non-initial mora had an H tone, while in an initial-accent word, the initial mora had an H tone, and a non-initial mora had an L tone. In the following, L or H will be used to note the tone of the target mora.

The target syllable was either a light syllable (CV) or a heavy syllable. A heavy syllable contained two moras: a long vowel or a diphthong vowel (CVV), a vowel closed by a plosive coda forming a geminate ('Q' in the usual Japanese notation) with the following onset (CVQ), or a vowel closed by a placeless nasal coda 'N' (CVN). Table 1 illustrates all the rimes used after target Cs in different conditions. The complete wordlist is given in Appendix 1.

Table 1 ***Rimes after target Cs, by syllable position, syllable structure, and tone. '#' stands for word boundary, and '.' for syllable boundary***

syllable position	S1												S2	
syllable structure	#CV(V).						#CVQ.				#CVN.		.CV(V/N)	
tone	L	H	L	H	L	H	L	H	L	H	L	H	L	H
pb-	aː	a	eː	e	–	i	ak	ak	–	–	eN	eN	a	ai
td-	ai	ai	e	e	–	–	ak	at	ep	ek	eN	eN	ai	ai
kg-	ai	ai	eː	eː	i	i	ak	at	ek	ek	eN	eN	a	ai
m-	aː	ai	–	–	–	–	–	–	–	–	–	–	aː	ai

In the first syllable (S1), three vowel contexts, /a, e, i/, were used after the plosive onsets. However, alveolar plosives are realized phonetically as alveolo-palatals in front of an /i/; thus no /t-di/ moras were included. The other gaps in Table 1 with /p, b/ onsets were due to the difficulty of finding a frequently used (near-)minimal pair. (Word-initial /p/ is almost always used in loan words.) In the second syllable (S2), as well as for the /m/ onset in all positions, only the vowel context /a/ was used.

We are aware of two shortcomings of in our speech materials. First, for open syllables CV(V), a mixture of heavy and light syllables should ideally have been avoided, since the pitch

[1] The selection of the speech materials was made with the great help of segmental neighborhoods calculated by Mafuyu Kitahara.

realization is more variable in an initial CVV than in an initial CV syllable (Kamiyama, 2003). This shortcoming was due to the difficulty of finding minimal pairs with solely heavy or light syllables. However, we ensured that each (near-)minimal pair contained the same structure in the target syllable. Second, the number of vowel contexts and syllable structures was much greater for S1 than S2, making the two positions somewhat unbalanced. This was because word-medial position has rarely been examined in previous studies; therefore, we intended to minimize the contextual variation by using the most neutral vowel, /a/, in this position. On the other hand, word-initial devoicing and aspiration have been reported previously; therefore, we intended to further investigate the role of vowel contexts and syllable structures.

Speakers were recorded individually in a sound-proof room with a Sony ECM-MS957 microphone through an Edirol Audio Interface connected to a laptop computer. A pop filter was placed between the microphone and the speaker to avoid strong aspiration noise that might cause audio clipping or low-frequency fluctuation in a waveform. The speaker first read each word in citation, and then the same words in the carrier sentence (2). The object particle "o" was preferred to the topic particle "wa" to avoid topicalization of "sore," which often introduces a pause after the particle. Nevertheless, even though "o" was used, a pause was quite often inserted between "o" and the target word (see Section 2.3 for details).

(2) sore o XX-to iu
That -OBJ. XX-COMP. say.
(I/We/People) call it XX.

All the tokens were presented on a laptop screen in a different random order for each speaker. Oral and written instructions were given in Japanese. Words having a Chinese written form in *kanji* (Chinese characters) were presented in both *kanji* and *katakana* (Japanese writing). Each token was repeated twice, for a total of 5616 tokens (78 words×2 contexts<citation/carrier sentence>×2 repetitions×18 speakers). The recording session also contained tokens with fricative onsets, but in this study, we shall focus exclusively on plosive onsets. Each recording session lasted approximately 45 minutes, including three or four short breaks. Participants then completed a short questionnaire about their basic personal information, linguistic experience, and music training background.

2.3 Positional and prosodic conditions

As explained in Section 2.2, target Cs can be in word-initial or word-medial positions. Moreover, we intend to compare words in citation form with words produced in the carrier sentence (2). However, a pause was frequently inserted after the particle "o" and before the target word, possibly due to a focalization process. Indeed, Beckman and Pierrehumbert (1986) reported that a pause often introduces an "intermediate-phrase" boundary in Japanese, right before a focused word or phrase.

Taking this into consideration, we defined five conditions in our study. For word-initial position: (i) WI_CITATION: word-initial position in citation form; (ii) WI_CARRIER-FOCUS: post-pausal word-initial position in a carrier sentence, considered as under focus; and (iii) WI_CARRIER: word-initial position in a carrier sentence without a preceding pause. For word-medial position: (iv) WM_CITATION: word-medial position in citation form and (v) WM_CARRIER: word-medial position in a carrier sentence. In addition, based on the presence or absence of a preceding pause, WI_CITATION and WI_CARRIER-FOCUS are defined as POST-PAUSAL, and WI_CARRIER, WM_CITATION, and WM_CARRIER are defined as INTERSONORANT. Table 2 indicates the number of analyzed tokens for each condition, with /t-d/ minimal pairs as examples.

It should be noted that the only indication we used to define a form under focus was the presence of a pause. Without further analyses of other acoustic effects related to prosodic boundaries, this definition should be taken as approximate. Moreover, a pause was not always easily distinguishable from the closure portion of a voiceless plosive. Before a voiceless plosive, we judged the presence of a pause on the basis of visualization of the waveform and the spectrogram, taking a combination of the following criteria into account: (a) the duration of the silent portion (more than 100 ms for most cases); (b) the presence of glottalization before the silence, although it was not obligatory; (c) the absence of visible formant transitions at the end

of the previous vowel. For 111 tokens, it was difficult to determine whether the silent portion was a pause or not; for these, the position was not included as a factor.

Table 2 ***The examined conditions of target Cs, and the number of analyzed tokens of each condition, with /t-d/ minimal pairs as examples. A vertical bar indicates a pause***

word-initial (n=4297)			word-medial (n=930)	
POST-PAUSAL (n=3245)		INTERSONORANT (n=1982)		
WI_CITATION (n=2230)	WI_CARRIER-FOCUS (n=1053)	WI_CARRIER (n=1082)	WM_CITATION (n=469)	WM_CARRIER (n=461)
taike:	sore o | **t**aike: to iu	sore o **t**aike: to iu	dʑi**t**ai	sore o dʑi**t**ai to iu
daike:	sore o | **d**aike: to iu	sore o **d**aike: to iu	dʑi**d**ai	sore o dʑi**d**ai to iu

2.4 Measurements and analyses

282 tokens (5% of the entire dataset), were produced with a pitch-accent pattern deviant from the prescriptions in the database, or with reading errors, and were excluded from all analyses. The data were segmented and annotated manually with Praat (Boersma & Weenink, 1992-2017). We measured voicing-related properties for voiced vs. voiceless plosives, which included VOT and voicing-ratio, as well as closure duration. We also measured *f*0 on the vowel following plosive and /m/ onsets.

Voice onset time (VOT). VOT was measured in POST-PAUSAL position (WI_CITATION and WI_CARRIER-FOCUS). For phonetically voiceless plosives (i.e., without voice lead), VOT intervals were marked between the onset of the release burst and the onset of the following vowel. The first ascending zero-crossing point on the waveform at F1 onset was determined as the vowel onset. In the syllable /ki/, followed by /s/ in our materials, the vowel /i/ was often devoiced in the L tone context (62 out of 72 tokens), and much less often in the H tone context (8 out of 72 tokens). (The devoicing of /i/ also occurred once in /gi/ followed by /s/.) VOT was not measurable for /k, g/ before a devoiced /i/. For phonetically voiced plosives (i.e., with voice lead), VOT intervals were measured between the onset of the release burst and the onset of regular glottal pulses.

Voicing-ratio. In INTERSONORANT position (WI_CARRIER, WM_CITATION, and WM_CARRIER), voicing-ratio (Snoeren et al., 2006) during the entire plosive, that is, closure plus release, was measured. This corresponded to the voiced duration relative to the duration of the entire plosive. Voicing-ratio values ranged from 0 (no voicing at all) to 1 (complete voicing). To ensure comparability of our results with previous studies (see Section 1.1.1), voicing-ratio was also measured during the closure portion of plosives.

Closure duration. In INTERSONORANT position, closure duration was measured for plosives. The start of the closure was defined as the moment of the offset of the preceding vowel, signaled by the disappearance of the high-frequency range formants (above F2) of the vowel on the spectrogram. The end of the closure was defined as the moment of the onset of the release burst of the plosive.

Fundamental frequency (f0). *f*0 was measured on the moraic vowel after target onsets. The first ascending zero-crossing point on the waveform at F1 onset was determined as the vowel onset, and the disappearance of the high-frequency range formants (above F2) as the vowel offset. The vowel was divided into 50 equal time intervals, and the mean *f*0 of each time interval was measured. In the case of diphthongs, the boundary between the two vowels being generally unclear, the first half of the diphthong was measured. Vowels followed by a nasal coda were generally easy to separate from the nasal coda by inspection of the waveform and the spectrogram. We visualized individual plots of raw *f*0 curves for each token by each speaker and excluded 91 tokens (i.e., 1.6% of the entire dataset), which manifested visible deviant *f*0 curves that were caused by detection errors (mainly octave errors) by Praat. The raw *f*0 values were then normalized using a within-speaker z-score transformation so that between-speaker variations were minimized. For analyses of the VOT-*f*0 correlation, in order to make the *f*0 difference more comparable between speakers, the *f*0 values were converted from hertz to semitones relative to each speaker's mean *f*0.

Statistical methods. Statistical models were constructed using functions in the *lme4* package (Bates, Maechler, Bolker, & Walker, 2017) in R (R Core Team, 2017). Generalized linear mixed models (GLMM) were built to fit to binary data using the *glmer* function. Linear mixed-effects models (LME) were built to fit to continuous data using the *lmer* function. All models included random intercepts for *speaker* and *item* and *by-speaker* random slopes for each predictor provided that they converged and improved the fitness of the model as evaluated by AIC. Predictors specified with levels were included as factor variables; otherwise, they were included as numerical variables. Likelihood-ratio comparisons, using the *anova* function in R, were conducted to assess the effect of each predictor and will be reported when relevant. Since we were especially interested in the contrasts between levels (e.g., between different conditions, places of articulation, and vowels) as well as the interactions between factors, we will mainly report post-hoc pairwise comparisons of the GLMM and LME models using the *emmeans* package (Lenth, 2018) in R, with *p*-values adjusted using the Tukey method. *Sex* and the interactions between *sex* and all the other factors were included as predictors in all the models, and all the results will be presented separately for male and female speakers. In some models, interactions involving more than two factors were included as predictors when the interaction was of interest and improved the fitness of the model.

3. Results

3.1 Voiced plosives: frequent devoicing in post-pausal position

We shall use "prevoiced" to describe plosives with a voice lead on the acoustic signal, and "devoiced" to describe phonologically voiced plosives without a voice lead on the acoustic signal.

In POST-PAUSAL position (WI_CITATION and WI_CARRIER-FOCUS), when VOT was negative, the token was counted as having a *prevoiced* plosive, including cases in which voicing was initiated but interrupted before the release. When VOT was zero or positive, the token was counted as having a *devoiced* plosive. Table 3 shows the mean VOT for prevoiced and devoiced plosives. In INTERSONORANT position (WI_CARRIER, WM_CITATION, and WM_CARRIER), we considered as *prevoiced* the plosives whose voicing-ratio was above 0.5, which is comparable to the suggestion of Abramson and Whalen (2017). 16.5% of the plosives were lenited in INTERSONORANT position, manifested by an absence of release. In this case, voicing-ratio was also measured throughout the consonant part. (Voicing-ratio was 0.16 on average (ranging from 0 to 0.68) for phonologically voiceless plosives in our data, resulting from the partial voicing that continues from the sonorant preceding the plosive.)

Table 3 ***VOTs (in ms, with standard deviations) of post-pausal prevoiced and devoiced plosives by place of articulation, position, and tone context***

		PREVOICED				DEVOICED			
		WI_CITATION		WI_CARRIER-FOCUS		WI_CITATION		WI_CARRIER-FOCUS	
	POA	L	H	L	H	L	H	L	H
Female	/b/	-60 (35)	-66 (33)	-66 (28)	-75 (27)	14 (5)	15 (5)	16 (7)	15 (6)
	/d/	-78 (29)	-71 (16)	-78 (28)	-79 (20)	15 (4)	14 (4)	15 (4)	14 (5)
	/g/	-72 (27)	-63 (29)	-95 (26)	-91 (24)	26 (9)	24 (5)	23 (7)	22 (7)
	mean	**-68**		**-81**		**18**		**18**	
Male	/b/	-66 (30)	-68 (31)	-105 (21)	-103 (29)	17 (6)	16 (5)	19 (5)	17 (3)
	/d/	-76 (26)	-70 (26)	-96 (20)	-102 (29)	18 (5)	18 (4)	19 (4)	14 (3)
	/g/	-68 (22)	-57 (22)	-121 (28)	-114 (28)	32 (11)	30 (11)	35 (11)	33 (8)
	mean	**-68**		**-107**		**22**		**23**	

Table 4 shows the percentage of prevoiced tokens in each positional/prosodic condition (henceforth "position"), separately for L and H tones, based on VOT for POST-PAUSAL position and on voicing-ratio for INTERSONORANT position. (The absence of /d/ data females for L tone in WM_CARRIER is because the word [sedai], noted with an initial accent (HLL) in the NTT database, was produced as accentless (LHH) by almost all female speakers in WM_CARRIER. After noticing this problem, we used another word pair, [ɕjotai] – [ɕjodai], for our later

recording with male speakers.) As shown in Table 4, in POST-PAUSAL position, less than half of the tokens by females and less than two-thirds of the tokens by males contained a prevoiced plosive onset, with a higher percentage of *prevoiced* realizations in WI_CARRIER-FOCUS than in WI_CITATION. In INTERSONORANT position, a great majority of phonologically voiced plosives were prevoiced, with a higher percentage in WM_CITATION and WM_CARRIER than in WI_CARRIER.

Table 4 ***Percentage of tokens with prevoiced plosives, by place of articulation, position, and tone context***

		POST-PAUSAL				INTERSONORANT					
		WI_CITATION		WI_CARRIER-FOCUS		WI_CARRIER		WM_CITATION		WM_CARRIER	
	POA	L	H	L	H	L	H	L	H	L	H
Female	/b/	27%	36%	44%	75%	80%	82%	94%	94%	94%	100%
	/d/	27%	30%	53%	70%	90%	94%	100%	89%	—	94%
	/g/	36%	42%	57%	61%	78%	84%	89%	89%	89%	83%
	mean	**34%**		**60%**		**84%**		**93%**		**92%**	
Male	/b/	46%	61%	56%	74%	77%	85%	82%	100%	89%	94%
	/d/	49%	60%	73%	83%	75%	94%	93%	78%	100%	69%
	/g/	36%	64%	76%	74%	69%	83%	78%	78%	100%	94%
	mean	**53%**		**73%**		**80%**		**85%**		**91%**	

With respect to the voicing pattern in POST-PAUSAL position, a large variability was observed both at the inter-speaker and within-speaker levels. Focusing on absolute initial position (WI_CITATION), the percentage of prevoiced plosives across all speakers ranged from 13% to 84% (Figure 1). It is also worth noting that for more than half of the speakers, the prevoiced percentage was lower than 50%. For individual speakers, it was not uncommon that the same speaker produced variable voicing for the same item in the same position. Figure 2 shows the waveforms and the spectrograms of two repetitions of the same word /dai.ke:/ produced by one female speaker, the first time with a prevoiced /d/, and the second time with a devoiced /d/. With respect to the voicing pattern in WI_CARRIER, we also measured voicing-ratio during closure, that is, the percentage of voiced duration relative to closure duration. It was at 80.8%, on average, for all the voiced plosives in WI_CARRIER, between the 57% found for English (Docherty, 1992) and the 94.6% found for Russian (Kulikov, 2012). We adopted the same criteria as Beckman et al. (2013) to assess whether Tokyo Japanese has a *passive voicing* pattern, as does German, or an *active voicing* pattern, as does Russian. The threshold of above 90% voicing-ratio during closure was used to define full voicing during closure. In WI_CARRIER, only 49.9% of voiced plosives were fully voiced. If the lenited realizations were added, this rate reached 53.6%. It was still lower than in German (62.5%), and obviously much lower than in Russian (97.5%).

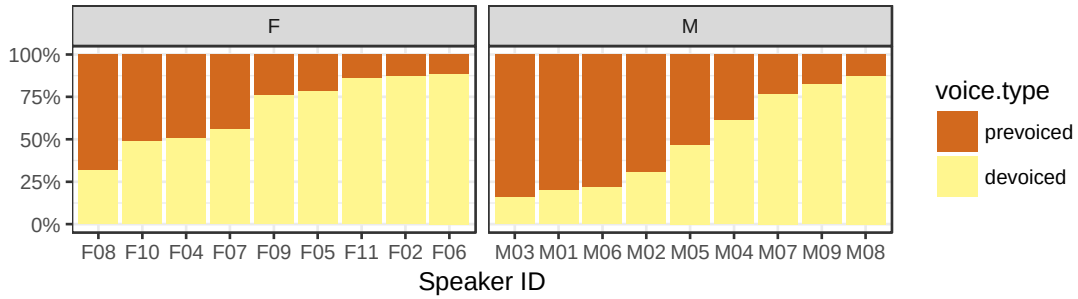


Figure 1 ***Percentage of prevoiced vs. devoiced plosives in WI_citation by speaker: left panel for females, and right panel for males***

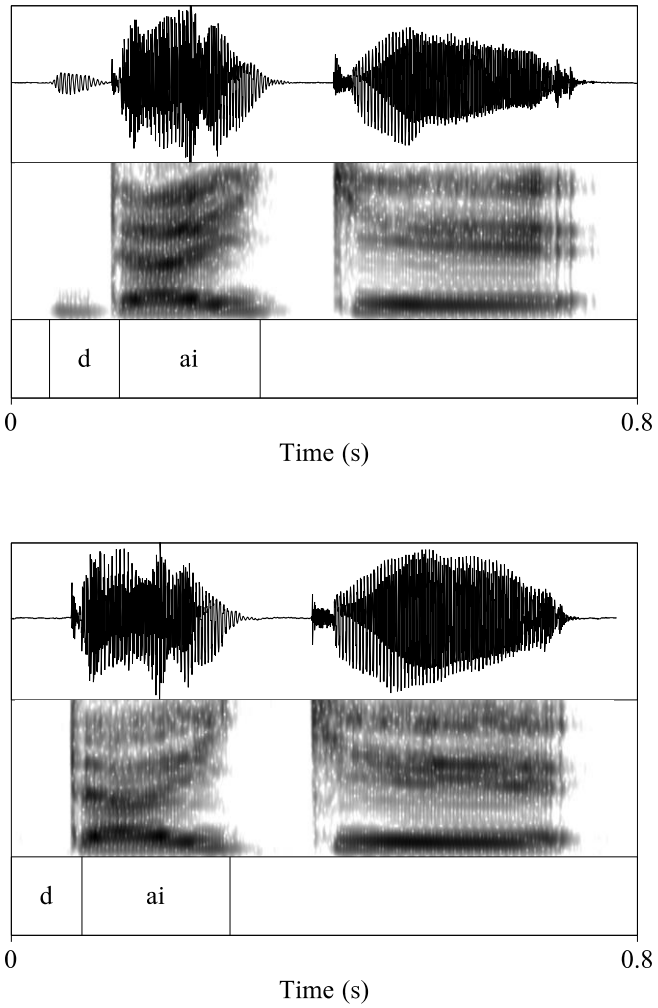


Figure 2 ***Within-speaker variation of voicing of*** **/d/** ***in the same word*** **/dai.ke:/** ***in WI_citation***

In order to assess the effects of various factors on the likelihood of voiced plosives being prevoiced, a GLMM model was selected to fit to the binary data (prevoiced or devoiced). The following predictors were included: *place of articulation* (POA: /b/, /d/, /g/); *vowel* (/i/, /e/, /a/); *syllable structure* (CV(V), CVQ, CVN); *position* (WI_CITATION, WI_CARRIER-FOCUS, WI_CARRIER, WM_CITATION, WM_CARRIER); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); a three-way interaction, *position* × *pitch-accent* × *sex*; as well as two-way interactions, *POA* × *sex*, *vowel* × *sex*, and *syllable structure* × *sex*. The model also included random intercepts for *speaker* and *item* as well as *by-speaker* random slopes for *position* and *POA*. The summary of the full model is given in Appendix 2 (MODEL 1). The results of the likelihood-ratio comparisons are shown in the brackets for main factors.

Place of articulation [$\chi^2 = 0.0005$, df = 2, $p = 1.00$]. The probability of prevoiced responses did not differ depending on POA. It has been suggested that labial plosives, because of the more expanded vocal tract, are more prone to voicing than alveolar and velar ones (Ohala & Riordan, 1979; Ohala, 1983). However, our Tokyo Japanese data do not provide evidence to confirm this tendency. *Vowel* [$\chi^2 = 25.96$, df = 2, $p < .001$]. Table 5 shows the results of post-hoc comparisons of log-odds for prevoiced realization among the three vowels. It has been suggested that voicing during closure of a plosive is better facilitated when it is followed by a high than a low vowel, because of the enlargement of the pharyngeal cavity during the production of a high vowel (Ohala & Riordan, 1979). Our results indeed showed an increasing percentage of prevoiced plosives in the following order: /a/ < /e/ < /i/ (56.8 < 68.9 < 73.8%), but the difference was significant only between the low /a/ and the non-low /i, e/ for females. (The

difference in estimated marginal means indicates the difference in the probability of prevoicing for each contrast, with higher values for higher probabilities.) Finally, *syllable structure* had no significant effect [$\chi^2 = 1.08$, df = 2, $p = .58$].

Table 5 ***Pairwise comparisons of prevoiced log-odds ratio among the three vowels, * for*** **p < *.05***

	Contrast	Estimate	SE	z.ratio	p.value
Female	i-e	0.46	0.32	1.44	0.323
	i-a	1.65	0.34	5.07	<.0001 *
	e-a	1.20	0.21	5.72	<.0001 *
Male	i-e	-0.29	0.32	-0.89	0.645
	i-a	0.13	0.32	0.42	0.910
	e-a	0.42	0.21	2.01	0.110

Position × *pitch-accent* × *sex* [$\chi^2 = 23.52$, df = 13, $p < .05$]. *Pitch-accent* [$\chi^2 = 8.40$, df = 1, $p = 0.004$]. The percentage of prevoiced plosives was higher in H than L tone contexts (67.7% > 57.6%). Table 6 shows the results of post-hoc comparisons of log-odds for prevoiced realizations between L and H tone contexts for each position, confirming a significant effect of *pitch-accent* in WI_CITATION and WI_CARRIER for males, and in WI_CARRIER-FOCUS for females. *Position* [$\chi^2 = 25.43$, df = 4, $p < .001$]. Table 7 shows the results of post-hoc comparisons of log-odds for prevoiced realizations among the five positions for each tone context. For female speakers, the probability of prevoicing of voiced plosives increases in the following order: for both H and L tone contexts, WI_CITATION < WI_CARRIER-FOCUS = WI_CARRIER, and WI_CARRIER-FOCUS < WM_CITATION = WM_CARRIER; and for the L tone context only, WI_CARRIER < WM_CITATION. In other words, the devoicing of /b, d, g/ is favored by word-initial position, and the devoicing tendency is stronger in the absolute initial position (WI_CITATION). For male speakers, the probability of prevoicing of voiced plosives increases in the following order: for the L tone context, WI_CITATION = WI_CARRIER-FOCUS < WM_CITATION, and WI_CARRIER < WM_CARRIER; and for the H tone context, WI_CARRIER-FOCUS < WI_CARRIER. In summary, there is a general trend for voiced plosives to be devoiced in domain-initial position (word-initial, especially POST-PAUSAL), but this trend is less clear-cut for male than for female speakers. Finally, the probability of prevoicing did not differ between males and females [$\chi^2 = 1.42$, df = 1, $p = .23$].

Table 6 ***Pairwise comparisons of prevoiced log-odds ratio between L and H tone contexts by position, * for*** **p < *.05***

	Position	Contrast	Estimate	SE	z.ratio	p.value
Female	WI_CITATION	L – H	-0.29	0.22	-1.30	0.193
	WI_CARRIER-FOCUS	L – H	-0.77	0.29	-2.70	0.007 *
	WI_CARRIER	L – H	-0.33	0.40	-0.81	0.419
	WM_CITATION	L – H	0.31	0.80	0.39	0.697
	WM_CARRIER	L – H	-0.05	0.82	-0.07	0.947
Male	WI_CITATION	L – H	-1.08	0.23	-4.69	<.0001 *
	WI_CARRIER-FOCUS	L – H	-0.56	0.36	-1.56	0.119
	WI_CARRIER	L – H	-1.24	0.35	-3.41	0.0005 *
	WM_CITATION	L – H	-0.21	0.63	-0.34	0.732
	WM_CARRIER	L – H	1.41	0.86	1.65	0.099

Table 7 ***Pairwise comparisons of prevoiced log-odds ratio among the positions by tone context, * for*** **p < *.05***

	Tone	Contrast	Estimate	SE	z.ratio	p.value
Female	L	WI_CITATION – WI_CARRIER-FOCUS	-1.18	0.43	-2.76	0.046 *
		WI_CITATION – WI_CARRIER	-2.46	0.78	-3.15	0.014 *
		WI_CITATION – WM_CITATION	-5.11	1.10	-4.65	<.0001 *
		WI_CITATION – WM_CARRIER	-4.77	0.94	-5.06	<.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	-1.28	0.72	-1.76	0.395
		WI_CARRIER-FOCUS – WM_CITATION	-3.93	1.00	-3.92	0.001 *
		WI_CARRIER-FOCUS – WM_CARRIER	-3.59	0.85	-4.22	0.0002 *

		WI_CARRIER – WM_CITATION	-2.65	0.95	-2.79	0.042 *
		WI_CARRIER – WM_CARRIER	-2.31	0.89	-2.60	0.070
		WM_CITATION – WM_CARRIER	0.34	1.03	0.33	0.998
	H	WI_CITATION – WI_CARRIER-FOCUS	-1.66	0.43	-3.86	0.001 *
		WI_CITATION – WI_CARRIER	-2.49	0.79	-3.17	0.013 *
		WI_CITATION – WM_CITATION	-4.51	1.01	-4.47	0.0001 *
		WI_CITATION – WM_CARRIER	-4.53	0.88	-5.15	<.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	-0.83	0.73	-1.13	0.789
		WI_CARRIER-FOCUS – WM_CITATION	-2.84	0.91	-3.14	0.015 *
		WI_CARRIER-FOCUS – WM_CARRIER	-2.87	0.78	-3.67	0.002 *
		WI_CARRIER – WM_CITATION	-2.01	0.85	-2.37	0.124
		WI_CARRIER – WM_CARRIER	-2.04	0.83	-2.47	0.097
		WM_CITATION – WM_CARRIER	-0.03	0.89	-0.03	1.000
Male	L	WI_CITATION – WI_CARRIER-FOCUS	-0.50	0.49	-1.04	0.837
		WI_CITATION – WI_CARRIER-FOCUS	-1.78	0.71	-2.51	0.088
		WI_CITATION – WM_CITATION	-2.99	0.97	-3.09	0.017 *
		WI_CITATION – WM_CARRIER	-4.38	1.01	-4.35	0.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	-1.28	0.67	-1.90	0.317
		WI_CARRIER-FOCUS – WM_CITATION	-2.48	0.90	-2.77	0.045 *
		WI_CARRIER-FOCUS – WM_CARRIER	-3.87	0.97	-3.98	0.0007 *
		WI_CARRIER – WM_CITATION	-1.20	0.72	-1.67	0.458
		WI_CARRIER – WM_CARRIER	-2.59	0.85	-3.03	0.021 *
		WM_CITATION – WM_CARRIER	-1.39	0.98	-1.43	0.611
	H	WI_CITATION – WI_CARRIER-FOCUS	0.01	0.48	0.03	1.000
		WI_CITATION – WI_CARRIER	-1.94	0.75	-2.61	0.069
		WI_CITATION – WM_CITATION	-2.12	0.97	-2.19	0.184
		WI_CITATION – WM_CARRIER	-1.88	0.80	-2.37	0.124
		WI_CARRIER-FOCUS – WI_CARRIER	-1.96	0.70	-2.79	0.043 *
		WI_CARRIER-FOCUS – WM_CITATION	-2.13	0.90	-2.37	0.123
		WI_CARRIER-FOCUS – WM_CARRIER	-1.90	0.75	-2.54	0.083
		WI_CARRIER – WM_CITATION	-0.17	0.76	-0.23	0.999
		WI_CARRIER – WM_CARRIER	0.06	0.63	0.10	1.000
		WM_CITATION – WM_CARRIER	0.24	0.77	0.31	0.998

3.2 Voiceless plosives: aspiration in word-initial position

Lisker and Abramson (1964) defined short-lag plosives as having a VOT of shorter than 30 ms and long-lag plosives as having a VOT of longer than 50 ms. As shown in Table 8, in word-initial position, phonologically voiceless plosives have, on average, a medium-lag VOT, that is, between the usual long-lag and short-lag VOTs, suggesting moderate aspiration. In word-medial position, they have a short-lag VOT, on average, suggesting no aspiration.

Table 8 ***VOTs (in ms, with standard deviations) of voiceless plosives by place of articulation, position, and tone context***

		POST-PAUSAL				INTERSONORANT					
		WI_CITATION		WI_CARRIER-FOCUS		WI_CARRIER		WM_CITATION		WM_CARRIER	
	POA	L	H	L	H	L	H	L	H	L	H
F	/p/	41 (17)	35 (14)	39 (16)	34 (13)	35 (18)	29 (11)	14 (5)	14 (5)	16 (6)	17 (7)
	/t/	37 (15)	34 (12)	35 (14)	32 (11)	25 (10)	24 (9)	13 (2)	16 (3)	14 (2)	15 (4)
	/k/	61 (17)	57 (16)	53 (13)	53 (13)	53 (14)	46 (11)	22 (8)	25 (8)	19 (5)	25 (7)
	mean	**44**		**41**		**35**		**17**		**18**	
M	/p/	34 (13)	34 (13)	39 (16)	41 (18)	33 (15)	33 (11)	19 (7)	16 (4)	17 (5)	17 (6)
	/t/	37 (12)	35 (13)	40 (18)	40 (18)	30 (7)	32 (11)	16 (4)	19 (5)	17 (4)	20 (5)
	/k/	56 (14)	56 (14)	64 (15)	64 (18)	51 (12)	50 (13)	25 (9)	28 (8)	27 (6)	30 (8)
	mean	**42**		**48**		**38**		**21**		**21**	

An LME model was selected to fit to the VOT data. The following predictors were included in the models: *place of articulation* (POA: /p/, /t/, /k/); *vowel* (/i/, /e/, /a/); *position* (WI_CITATION, WI_CARRIER-FOCUS, WI_CARRIER, WM_CITATION, WM_CARRIER); *pitch-accent* (L vs. H tone

contexts); *sex* (M vs. F); a three-way interaction, *position* × *pitch-accent* × *sex*; as well as two-way interactions, *POA* × *sex* and *vowel* × *sex*. (The model did not converge when *syllable structure* was added as a predictor.) The model also included random intercepts for *speaker* and *item* as well as *by-speaker* random slopes for *position*. The summary of the full model is given in Appendix 2 (MODEL 2). The results of the likelihood-ratio comparisons are shown in the brackets for main factors.

Place of articulation [$\chi^2 = 45.55$, df = 2, $p < .001$]. Table 9 shows the results of post-hoc comparisons of VOTs among the three POAs. For both males and females, /k/ has significantly longer VOTs either both /p/ or /t/, while /p/ does not differ from /t/. Indeed, it is attested in many languages that velar plosives have longer VOTs than labial and alveolar ones, for diverse physiological and aerodynamic reasons (for a review, see Cho & Ladefoged, 1999). *Vowel* [$\chi^2 = 24.33$, df = 2, $p < .001$]. Table 10 shows the results of post-hoc comparisons of VOTs among the three vowels. It is shown that in some languages, VOTs are lengthened when the voiceless plosive is followed by a high compared to a non-high vowel, although contradictory results have been reported (Nearey & Rochet, 1994). Our results showed an increasing VOT in the following order: /e/ < /a/ < /i/ (33 < 44 < 53 ms). The difference between all vowel pairs but /e/ and /a/ for males was significant.

Table 9 ***Pairwise comparisons of VOTs among the three POAs, * for* p < .05**

	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	p-t	1.75	2.17	52.44	0.81	0.697
	p-k	-18.37	2.11	51.03	-8.69	<.0001 *
	t-k	-20.14	2.07	62.41	-9.74	<.0001 *
Male	p-t	-2.40	2.16	52.41	-1.11	0.511
	p-k	-20.75	2.12	51.88	-9.77	<.0001 *
	t-k	-18.35	2.06	52.21	-8.89	<.0001 *

Table 10 ***Pairwise comparisons of VOTs among the three vowels, * for* p < .05**

	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	i-e	17.77	3.51	56.66	5.07	<.0001 *
	i-a	10.68	3.58	56.28	2.98	0.011 *
	e-a	-7.09	1.99	50.74	-3.56	0.002 *
Male	i-e	16.73	3.58	61.73	4.67	<.0001 *
	i-a	12.95	3.65	61.04	3.55	0.002 *
	e-a	-3.78	2.00	51.39	-1.89	0.151

Position × *pitch-accent* × *sex* [$\chi^2 = 36.30$, df = 13, $p < .001$]. *Pitch-accent* [$\chi^2 = 1.34$, df = 1, $p = .25$]. Table 11 shows the results of post-hoc comparisons of VOT between L and H tone contexts for each position. VOTs are significantly longer in L than in H tone contexts only for word-initial position for females. *Position* [$\chi^2 = 24.74$, df = 4, $p < .001$]. Table 12 shows the results of post-hoc comparisons of VOT among the five positions. For both males and females, VOTs are significantly longer in all word-initial than in all word-medial positions, while no difference is found between other positional or prosodic conditions. *Sex* [$\chi^2 = 4.07$, df = 1, $p < .05$]. The results of post-hoc comparisons of VOT between males and females (Table 13) show that VOTs are longer for males than females only in word-medial position (except WM_CITATION in the H context).

Table 11 ***Pairwise comparisons of VOT between L and H tone contexts by position, * for* p < .05**

	Position	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	WI_CITATION	L – H	6.09	1.97	62.97	3.09	0.003 *
	WI_CARRIER-FOCUS	L – H	5.22	2.10	84.81	2.48	0.015 *
	WI_CARRIER	L – H	6.91	2.21	106.56	3.12	0.002 *
	WM_CITATION	L – H	-2.89	4.24	76.06	-0.68	0.497
	WM_CARRIER	L – H	-2.44	4.20	73.03	-0.58	0.563
Male	WI_CITATION	L – H	2.87	1.98	64.46	1.45	0.153
	WI_CARRIER-FOCUS	L – H	1.47	2.27	118.52	0.65	0.520

	WI_CARRIER	L – H	0.91	2.13	89.79	0.43	0.671
	WM_CITATION	L – H	-1.49	4.14	69.39	-0.36	0.720
	WM_CARRIER	L – H	-0.68	4.14	69.50	-0.16	0.871

Table 12 ***Pairwise comparisons of VOT among the five positions by tone context, * for*** **p < *.05***

	Tone	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	L	WI_CITATION – WI_CARRIER-FOCUS	5.02	2.77	19.37	1.81	0.395
		WI_CITATION – WI_CARRIER	6.38	2.82	16.81	2.26	0.205
		WI_CITATION – WM_CITATION	34.72	4.16	69.43	8.34	<.0001 *
		WI_CITATION – WM_CARRIER	34.85	4.03	77.48	8.64	<.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	1.36	3.35	5.43	0.41	0.993
		WI_CARRIER-FOCUS – WM_CITATION	29.69	4.81	50.96	6.17	<.0001 *
		WI_CARRIER-FOCUS – WM_CARRIER	29.82	4.53	59.90	6.59	<.0001 *
		WI_CARRIER – WM_CITATION	28.34	4.99	54.47	5.68	<.0001 *
		WI_CARRIER – WM_CARRIER	28.47	4.72	62.27	6.04	<.0001 *
		WM_CITATION – WM_CARRIER	0.13	2.10	74.97	0.06	1.000
	H	WI_CITATION – WI_CARRIER-FOCUS	4.15	2.76	18.55	1.51	0.572
		WI_CITATION – WI_CARRIER	7.20	2.73	15.40	2.64	0.111
		WI_CITATION – WM_CITATION	26.19	4.38	66.23	5.98	<.0001 *
		WI_CITATION – WM_CARRIER	25.87	4.22	68.26	6.13	<.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	3.05	3.26	4.47	0.94	0.871
		WI_CARRIER-FOCUS – WM_CITATION	22.03	5.00	52.30	4.41	0.0005 *
		WI_CARRIER-FOCUS – WM_CARRIER	21.72	4.69	57.38	4.63	0.0002 *
		WI_CARRIER – WM_CITATION	18.98	5.13	54.99	3.70	0.004 *
		WI_CARRIER – WM_CARRIER	18.67	4.84	59.29	3.86	0.003 *
		WM_CITATION – WM_CARRIER	-0.31	1.86	49.45	-0.17	1.000
Male	L	WI_CITATION – WI_CARRIER-FOCUS	0.87	3.04	20.55	0.29	0.998
		WI_CITATION – WI_CARRIER	0.06	2.20	21.67	0.03	1.000
		WI_CITATION – WM_CITATION	24.43	4.09	65.54	5.97	<.0001 *
		WI_CITATION – WM_CARRIER	24.27	3.92	70.01	6.19	<.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	-0.81	2.70	11.03	-0.30	0.998
		WI_CARRIER-FOCUS – WM_CITATION	23.55	4.90	48.90	4.81	0.0001 *
		WI_CARRIER-FOCUS – WM_CARRIER	23.40	4.58	54.06	5.11	<.0001 *
		WI_CARRIER – WM_CITATION	24.37	4.60	52.24	5.29	<.0001 *
		WI_CARRIER – WM_CARRIER	24.21	4.27	59.72	5.67	<.0001 *
		WM_CITATION – WM_CARRIER	-0.16	1.93	55.04	-0.08	1.000
	H	WI_CITATION – WI_CARRIER-FOCUS	-0.53	2.97	18.70	-0.18	1.000
		WI_CITATION – WI_CARRIER	-1.90	2.20	21.60	-0.86	0.908
		WI_CITATION – WM_CITATION	20.88	4.39	66.83	4.76	0.0001 *
		WI_CITATION – WM_CARRIER	19.91	4.23	68.92	4.71	0.0001 *
		WI_CARRIER-FOCUS – WI_CARRIER	-1.37	2.64	9.54	-0.52	0.983
		WI_CARRIER-FOCUS – WM_CITATION	21.41	5.12	52.79	4.18	0.001 *
		WI_CARRIER-FOCUS – WM_CARRIER	20.44	4.82	57.45	4.24	0.0008 *
		WI_CARRIER – WM_CITATION	22.78	4.88	56.90	4.67	0.0002 *
		WI_CARRIER – WM_CARRIER	21.81	4.56	62.83	4.78	0.0001 *
		WM_CITATION – WM_CARRIER	-0.97	1.90	52.92	-0.51	0.986

Table 13 ***Pairwise comparisons of VOT between females and males by tone context and position, * for*** **p < *.05***

Tone	Position	Contrast	Estimate	SE	Df	t.ratio	p.value
L	WI_CITATION	female – male	3.89	4.10	21.75	0.94	0.356
	WI_CARRIER-FOCUS	female – male	-0.28	5.18	19.64	-0.06	0.957
	WI_CARRIER	female – male	-2.45	5.01	21.60	-0.49	0.630
	WM_CITATION	female – male	-6.42	2.65	50.35	-2.42	0.019 *
	WM_CARRIER	female – male	-6.71	2.61	54.50	-2.57	0.013 *
H	WI_CITATION	female – male	0.65	4.07	21.06	0.16	0.875
	WI_CARRIER-FOCUS	female – male	-4.03	5.12	18.64	-0.79	0.440
	WI_CARRIER	female – male	-8.45	4.95	20.79	-1.71	0.103
	WM_CITATION	female – male	-4.66	2.52	41.37	-1.85	0.072
	WM_CARRIER	female – male	-5.31	2.43	41.76	-2.19	0.035 *

3.3 Distinctiveness of the two plosive series

In POST-PAUSAL position (WI_CITATION and WI_CARRIER-FOCUS), the frequent devoicing of voiced plosives (see Section 3.1) led to an overlap of VOT between devoiced /b, d, g/ (short-lag on average) and voiceless plosives /p, t, k/ (medium-lag on average). In Figure 3, the distribution of VOT of the two plosive series is plotted for WI_CITATION by place of articulation. (A similar pattern can be found for WI_CARRIER-FOCUS.) For each POA, the overlap of VOT is visible, but the two categories are clearly separable. An LME model was selected to fit to the VOT data for devoiced /b, d, g/ and phonologically voiceless /p, t, k/ in POST-PAUSAL position. The following predictors were included: *phonological voicing* (voiced vs. voiceless); *POA* (labial, dental, velar); *vowel* (/i/, /e/, /a/); *syllable structure* (CV(V), CVQ, CVN); *position* (WI_CITATION vs. WI_CARRIER-FOCUS); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); a four-way interaction, *phonological voicing* × *position* × *POA* × *sex*; as well as the interactions between *sex* and the other two predictors. The model included random intercepts for *speaker* and *item*, as well as random *by-speaker* slopes for all the predictors. The summary of the full model is given in Appendix 2 (MODEL 3). The results of the likelihood-ratio comparisons showed the following effects for the following predictors: *phonological voicing* [$\chi^2 = 3.67$, df = 1, $p = .06$], and the *phonological voicing* × *position* × *POA* × *sex* interaction [$\chi^2 = 55.87$, df = 18, $p < .001$]. As shown by the post-hoc pairwise comparisons (Table 14), the VOT difference between the two (phonetically voiceless) plosive series is significant for each POA in each position for both males and females.

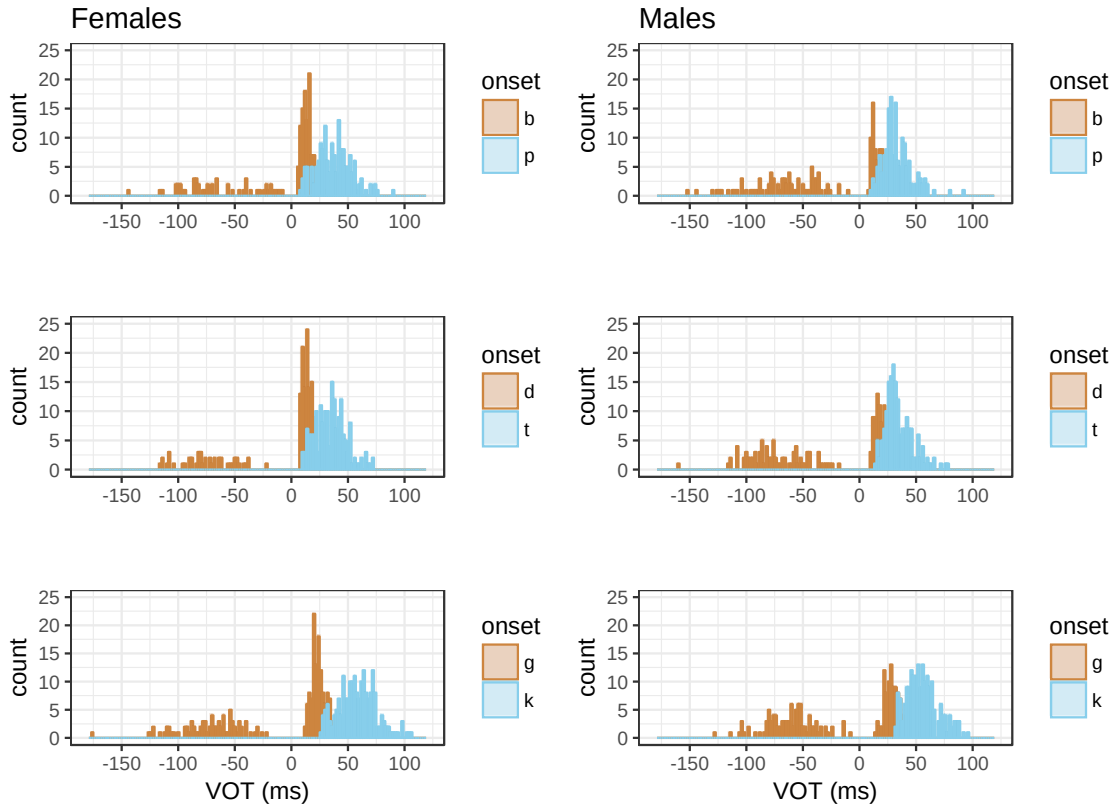


Figure 3 ***VOT distribution of the two plosive series by place of articulation in WI_citation***

Table 14 ***Pairwise comparisons of VOT between /p, t, k/ and devoiced /b, d, g/ by POA for each post-pausal position***

	Position	POA	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	WI_CITATION	labial	voiceless – devoiced	24.00	3.07	45.65	7.81	<.0001
		dental	voiceless – devoiced	21.09	3.02	42.87	6.99	<.0001
		velar	voiceless – devoiced	34.47	2.94	39.24	11.74	<.0001
	WI_CARRIER-	labial	voiceless – devoiced	20.74	3.40	69.51	6.10	<.0001

	FOCUS							
		dental	voiceless – devoiced	17.75	3.39	67.64	5.23	<.0001
		velar	voiceless – devoiced	30.20	3.19	54.61	9.48	<.0001
Male	WI_CITATION	labial	voiceless – devoiced	18.12	3.16	51.01	5.73	<.0001
		dental	voiceless – devoiced	18.19	3.11	47.81	5.85	<.0001
		velar	voiceless – devoiced	26.56	2.98	41.43	8.92	<.0001
	WI_CARRIER-FOCUS	labial	voiceless – devoiced	21.64	3.74	101.27	5.78	<.0001
		dental	voiceless – devoiced	21.51	4.00	135.07	5.38	<.0001
		velar	voiceless – devoiced	28.85	3.68	98.51	7.83	<.0001

In INTERSONORANT position (WI_CARRIER, WM_CITATION, and WM_CARRIER), prevoicing was frequently realized (see Section 3.1). Voicing-ratio was much higher for phonologically voiced plosives than for their voiceless counterparts. Besides, closure duration is often cited as an important cue signaling voicing (Lisker, 1957, 1986). This pattern was also found in our data, with longer closure duration for voiceless plosives than for their voiced counterparts. The average difference in closure duration between the two plosive series was, however, much larger in WM_CITATION (61 ms) and WM_CARRIER (40 ms) than in WI_CARRIER (10 ms). Figure 4 shows voicing-ratio and closure duration for each onset in INTERSONORANT position. (Lenited realizations were excluded from closure duration measures.) Two separate LME models were selected to fit to the voicing-ratio and closure duration data, respectively. The model for voicing-ratio data included random intercepts for *speaker* and *item*, as well as *by-speaker* random slopes for *pitch-accent*. The following predictors were included: *phonological voicing* (voiced vs. voiceless); *POA* (labial, dental, velar); *position* (WI_CARRIER, WM_CITATION, WM_CARRIER); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); as well as a three-way interaction, *phonological voicing* × *position* × *sex*; and a two-way interaction *sex* × *pitch-accent*. The model for closure duration data included random intercepts for *speaker* and *item*, as well as *by-speaker* random slopes for *pitch-accent* and *position.* The following predictors were included: *phonological voicing* (voiced vs. voiceless); *POA* (labial, dental, velar); *position* (WI_CARRIER, WM_CITATION, WM_CARRIER); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); as well as a three-way interaction *phonological voicing* × *position* × *sex*; and two-way interactions, *sex* × *pitch-accent* and *sex* × *vowel*. The summary of the full model is given in Appendix 2 (MODELS 4 and 5). The results of the likelihood-ratio comparisons showed the effects of *phonological voicing* in the two models: for voicing-ratio [χ^2 = 203.78, df = 1, $p < .001$], and for closure duration [χ^2 = 83.13, df = 1, $p < .001$]. As shown by the post-hoc pairwise comparisons of these two models (Tables 15 and 16), both the voicing-ratio and closure duration differences between the two plosive series are significant for each position for both males and females.

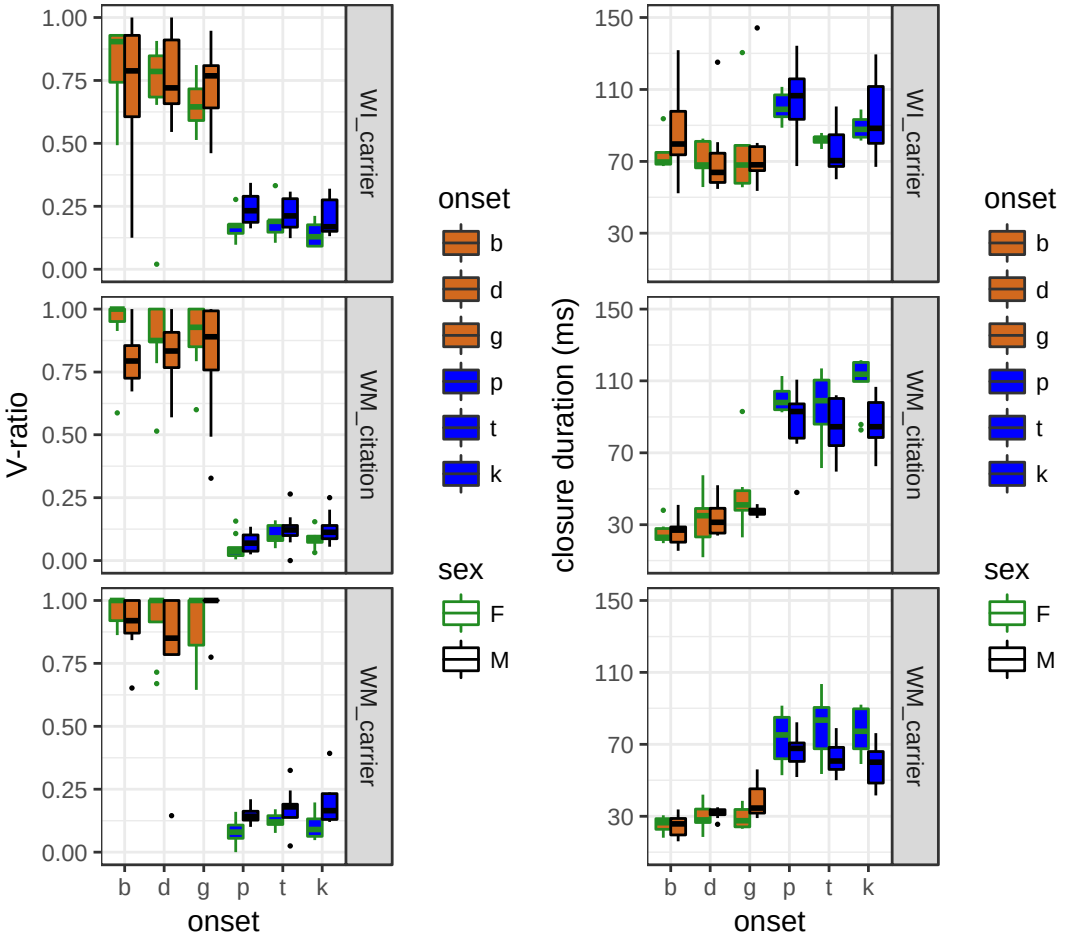


Figure 4 ***Boxplots of voicing-ratio (left panels) and closure duration (right panels) of intersonorant (WI_carrier, WM_citation, and WM_carrier) plosives by onset and position***

Table 15 Pairwise comparisons of voicing-ratio between the two plosive series for each intersonorant position

	Position	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	WI_CARRIER	voiceless – voiced	-0.58	0.02	152.28	-25.06	<.0001
	WM_CITATION	voiceless – voiced	-0.83	0.04	134.95	-18.87	<.0001
	WM_CARRIER	voiceless – voiced	-0.82	0.04	129.05	-18.69	<.0001
Male	WI_CARRIER	voiceless – voiced	-0.55	0.02	124.24	-24.84	<.0001
	WM_CITATION	voiceless – voiced	-0.71	0.04	114.25	-16.71	<.0001
	WM_CARRIER	voiceless – voiced	-0.74	0.04	114.59	-17.33	<.0001

Table 16 Pairwise comparisons of closure duration between the two plosive series for each intersonorant position

	Position	Contrast	Estimate	SE	Df	t.ratio	p.value
Female	WI_CARRIER	voiceless – voiced	19.14	1.79	260.77	10.68	<.0001
	WM_CITATION	voiceless – voiced	67.84	3.59	222.09	18.89	<.0001
	WM_CARRIER	voiceless – voiced	48.57	3.79	262.74	12.81	<.0001
Male	WI_CARRIER	voiceless – voiced	18.60	1.65	187.13	11.24	<.0001
	WM_CITATION	voiceless – voiced	53.51	3.35	203.07	15.97	<.0001
	WM_CARRIER	voiceless – voiced	32.55	3.58	244.10	9.09	<.0001

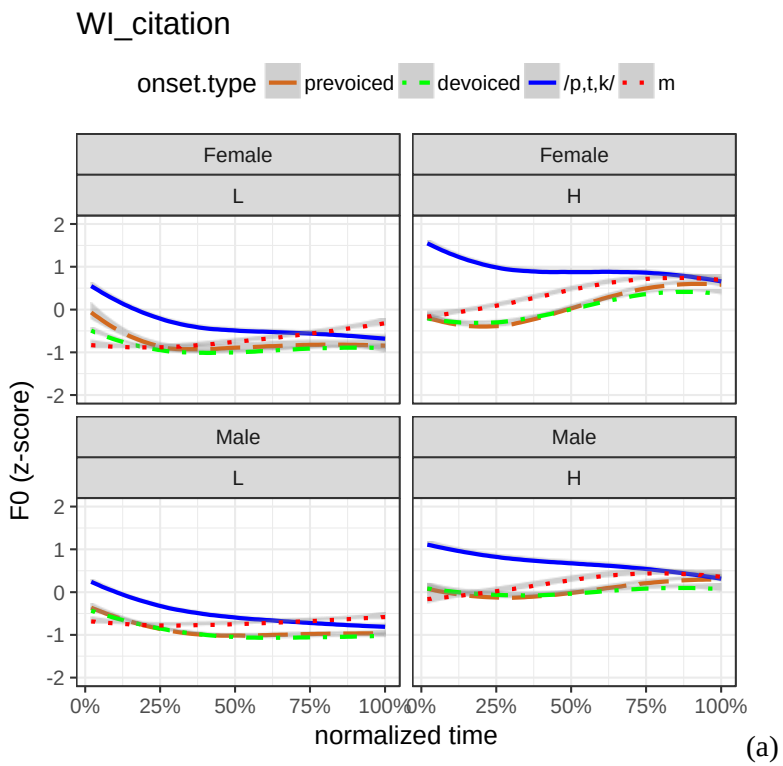
3.4 f0 enhancements in word-initial position

This section will present the *f*0 results on the vowels following plosives and /m/. Only the vowel context /a/ was used to compare plosives with /m/ onset. Time was normalized to show the duration of the *f*0 perturbations relative to the duration of the vowel. For an indication of the

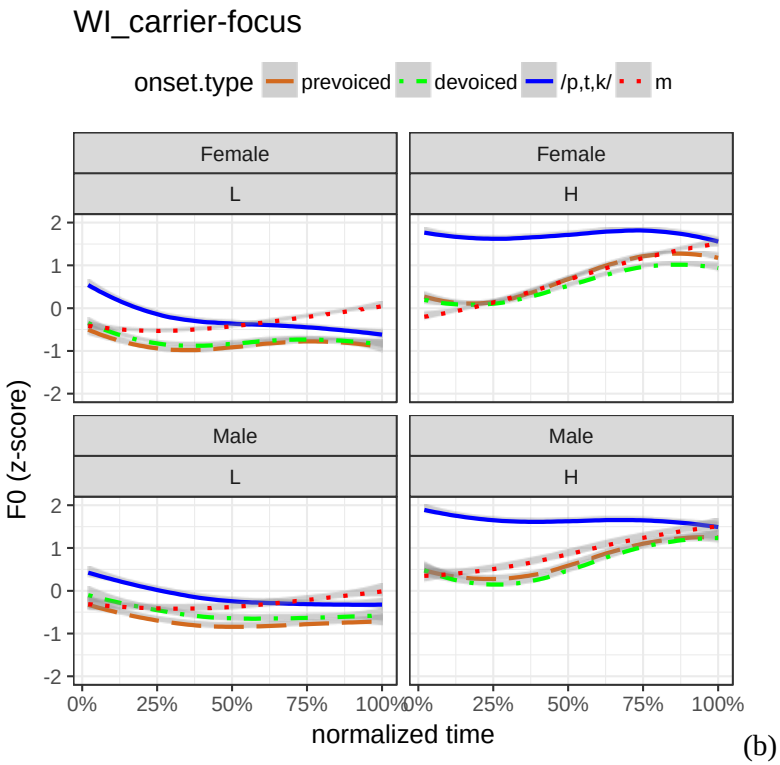
absolute duration of *f*0 perturbations, the mean moraic vowel duration across all contexts and positions was 112 ms for females and 91 ms for males. Figure 5 shows the time-normalized z-score *f*0 curves aggregated over speakers and items, following each onset type in each position but WM_CITATION. (In WM_CITATION, the target mora was produced in absolute final position. This led to strong creakiness and generated many chaotic *f*0 curves.) The curves were smoothed using LOESS fitting with 95% confidence intervals indicated by shading.

From Figure 5, we observe that the three word-initial positions pattern together. After word-initial onsets (Figure 5abc), *f*0 is higher after voiceless plosives than after /m/ or voiced plosives, whether phonetically prevoiced or devoiced. This *f*0 difference is larger for H than L tone contexts. For an indication of the difference in Hertz, the *f*0 curves in Hertz after phonologically voiceless and voiced plosives in word-initial position aggregated over speakers and items are given in Appendix 3, with all vowels included. For female speakers, *f*0 at vowel onset is higher following voiceless than voiced plosives by about 20 Hz in the L tone context, and about 40 Hz in the H tone context. The *f*0 difference is maintained until the vowel offset in the H tone context, and until about the middle of the vowel in the L tone context. For male speakers, the *f*0 difference at vowel onset is about 10 Hz in the H tone context and under 5 Hz in the L tone context. Besides, the *f*0 difference also lasts longer in H than L tone contexts. The individual raw *f*0 plots for WI_CITATION are given in Appendix 4. A more consistent *f*0 difference can be observed for female speakers than for male speakers.

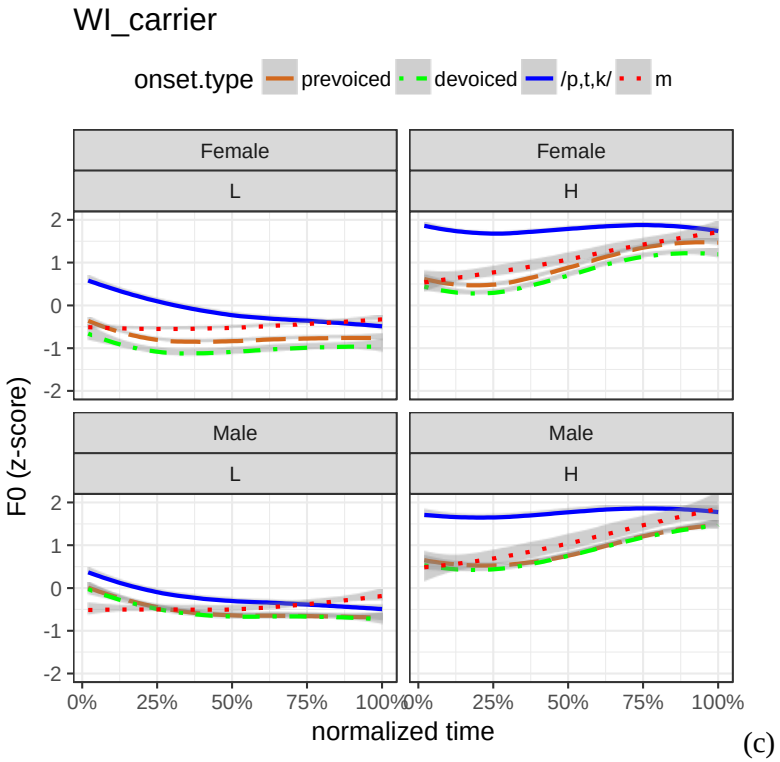
After word-medial onsets (Figure 5d), the *f*0 difference between onset types is much smaller than in word-initial position. (Tokens with lenited plosives were excluded.) The curves are very close to each other and the confidence intervals overlap here and there. In words with an initial accent, the first mora carries an H tone and the following moras carry an L tone. However, since the first syllable may contain either one or two moras, the *f*0 onset of the target mora in the second syllable was affected by coarticulation with the previous mora: the *f*0 onset was high when the previous mora carried an H tone, and low when the previous mora carried an L tone. We thus separated the L tone context further into L-low (for /t-d, k-g/) and L-high (for /p-b, m/) realizations.



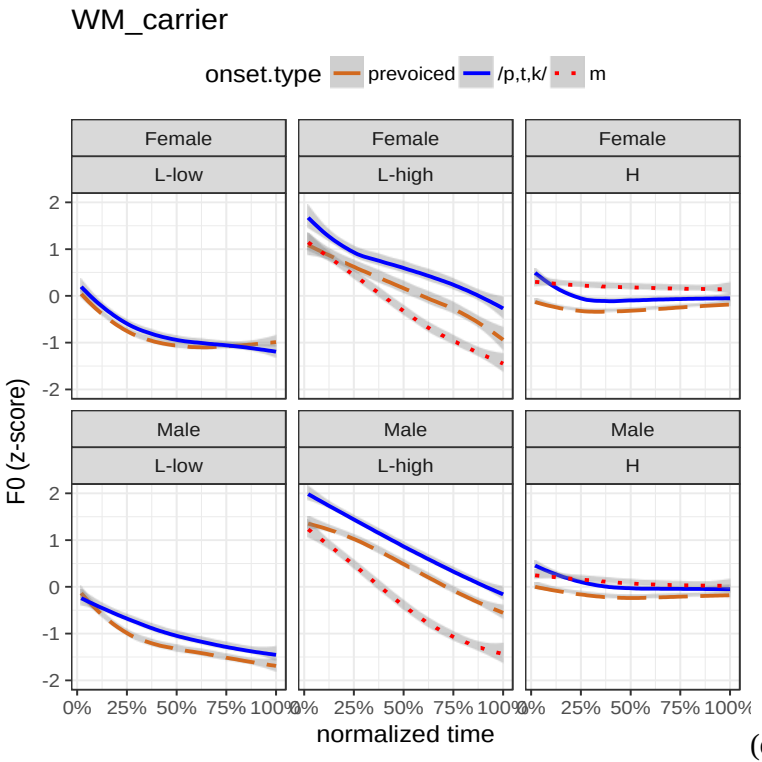
(a)


WI_carrier-focus
onset.type
prevoiced
devoiced
/p,t,k/
m
Female
L
Female
H
Male
L
Male
H
F0 (z-score)
2
1
0
-1
-2
0%
25%
50%
75%
100%
normalized time

(b)


WI_carrier
onset.type
prevoiced
devoiced
/p,t,k/
m
Female
L
Female
H
Male
L
Male
H
F0 (z-score)
2
1
0
-1
-2
0%
25%
50%
75%
100%
normalized time

(c)



(d)

Figure 5 ***Time- and z-score normalized f0 curves after plosives and /m/ by position, smoothed with LOESS fitting, with 95% confidence intervals indicated by shading***

Figure 5 clearly shows that word-initial positions (WI_CITATION, WI_CARRIER-FOCUS, and WI_CARRIER) pattern together against WM_CARRIER in terms of the *f*0 curves after each onset type. Two separate LME models were thus selected to fit to the z-score *f*0 data at the vowel onset (first 10% of the vowel): one for word-initial position, and the other for WM_CARRIER. We also attempted to fit a unique model integrating all positions but it failed to converge. The model for word-initial positions included the following predictors: *onset type* (/p, t, k/, devoiced, prevoiced, /m/); *position* (WI_CITATION, WI_CARRIER-FOCUS, WI_CARRIER); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); *time point*; and a four-way interaction, *onset type* × *position* × *pitch-accent* × *sex*. The interaction, which improved the fitness of the model, was included to examine how onset *f*0 differed between each pair of onset types, and how it interacted with the other factors. The model also included intercepts for *speaker* and *item*, as well as *by-speaker* random slopes for *onset type* and *position*. The summary of the full model is given in Appendix 2 (MODEL 6). The results of the likelihood-ratio comparisons showed the effect of *onset type* on onset *f*0 [$\chi^2 = 22.90$, df = 3, $p < .001$]. Post-hoc pairwise comparisons (Table 17) indeed confirmed that *f*0 at the vowel onset was higher after /p, t, k/ than /b, d, g, m/ in word-initial position, regardless of the presence or absence of closure voicing of /b, d, g/. This pattern was consistent across positions and tones (except for males in the L tone context in WI_CARRIER). In contrast, *f*0 did not differ after devoiced and prevoiced plosives (again, except for males in the L tone context in WI_CARRIER).

The model for WM_CARRIER included the following predictors: *onset type* (/p, t, k/, devoiced, prevoiced, /m/); *pitch-accent* (L vs. H tone contexts); *sex* (M vs. F); *time point*; as well as a three-way interaction, *onset type* × *pitch-accent* × *sex*. The model also included intercepts for *speaker* and *item*, as well as *by-speaker* random slopes for *onset type*. In WM_CARRIER, *onset type* had no effect on onset *f*0 [$\chi^2 = 1.07$, df = 3, $p = .79$]. There were few occurrences of devoiced plosives in this position, and the *f*0 pattern differed very occasionally without systematicity between devoiced and prevoiced plosives. Thus, for the sake of simplicity, Table 18 shows the results of the post-hoc pairwise comparisons only between /p, t, k/ and prevoiced plosives, confirming that onset *f*0 did not differ after these two plosive series.

Table 17 ***Pairwise comparisons of z-score f0 after word-initial onsets (over the first 10% of the vowel) among the four onset types by tone context, * for*** **p < .01**

		WI_CITATION									
	Tone	L					H				
	Contrast	*β*	SE	df	t.ratio	p.value	*β*	SE	df	t.ratio	p.value
F	p,t,k – dev.	1.07	0.14	47.94	7.61	<.0001*	1.79	0.14	48.44	12.70	<.0001*
	p,t,k – vcd.	0.89	0.14	65.99	6.03	0.0001*	1.89	0.14	60.29	13.16	<.0001*
	p,t,k – m	1.40	0.23	53.98	6.18	<.0001*	1.68	0.23	53.45	7.46	<.0001*
	dev. – vcd.	-0.18	0.09	59.77	-1.95	0.220	0.11	0.09	48.61	1.22	0.618
	dev. – m	0.33	0.20	46.85	1.67	0.358	-0.10	0.20	46.43	-0.53	0.952
	vcd. – m	0.51	0.21	54.72	2.47	0.076	-0.21	0.20	51.47	-1.04	0.726
M	p,t,k – dev.	0.70	0.14	53.28	4.83	0.0001*	1.06	0.14	53.57	7.31	<.0001*
	p,t,k – vcd.	0.64	0.14	57.64	4.48	0.0002*	1.16	0.14	53.39	8.29	<.0001*
	p,t,k – m	0.86	0.23	54.06	3.81	0.002*	1.31	0.23	53.48	5.81	<.0001*
	dev. – vcd.	-0.06	0.09	52.81	-0.67	0.909	0.10	0.09	45.90	1.19	0.638
	dev. – m	0.16	0.20	49.20	0.82	0.846	0.26	0.20	49.07	1.28	0.583
	vcd. – m	0.22	0.20	50.39	1.11	0.688	0.15	0.20	48.19	0.77	0.867
		WI_CARRIER-FOCUS									
	Tone	L					H				
	Contrast	*β*	SE	df	t.ratio	p.value	*β*	SE	df	t.ratio	p.value
F	p,t,k – dev.	0.81	0.15	55.60	5.60	<.0001*	1.61	0.15	59.58	10.89	<.0001*
	p,t,k – vcd.	0.93	0.15	64.89	6.29	<.0001*	1.55	0.15	59.39	10.66	<.0001*
	p,t,k – m	0.78	0.23	59.28	3.38	0.007*	1.94	0.23	59.35	8.39	<.0001*
	dev. – vcd.	0.11	0.10	63.07	1.17	0.647	-0.06	0.10	59.16	-0.63	0.921*
	dev. – m	-0.03	0.21	54.23	-0.17	0.998	0.33	0.21	56.32	1.59	0.392
	vcd. – m	-0.15	0.21	57.39	-0.70	0.898	0.39	0.21	54.93	1.88	0.249
M	p,t,k – dev.	0.71	0.17	98.65	4.27	0.0003*	1.64	0.18	110.54	9.17	<.0001*
	p,t,k – vcd.	0.88	0.15	77.15	5.76	<.0001*	1.71	0.15	68.84	11.46	<.0001*
	p,t,k – m	0.82	0.24	67.72	3.43	0.006*	1.74	0.24	66.40	7.32	<.0001*
	dev. – vcd.	0.17	0.13	173.32	1.35	0.533	0.07	0.14	166.92	0.49	0.962
	dev. – m	0.10	0.22	78.29	0.46	0.967	0.09	0.23	87.85	0.40	0.978
	vcd. – m	-0.07	0.22	66.01	-0.31	0.990	0.03	0.22	63.79	0.13	0.999
		WI_CARRIER									
	Tone	L					H				
	Contrast	*β*	SE	df	t.ratio	p.value	*β*	SE	df	t.ratio	p.value
F	p,t,k – dev.	1.41	0.16	90.91	8.66	<.0001*	1.71	0.16	83.71	10.74	<.0001*
	p,t,k – vcd.	1.09	0.15	65.66	7.30	<.0001*	1.47	0.15	63.05	9.98	<.0001*
	p,t,k – m	1.30	0.24	67.93	5.44	<.0001*	1.47	0.24	67.64	6.16	<.0001*
	dev. – vcd.	-0.32	0.12	147.48	-2.73	0.036	-0.24	0.11	118.51	-2.10	0.158
	dev. – m	-0.11	0.22	79.16	-0.51	0.958	-0.25	0.22	76.37	-1.11	0.686
	vcd. – m	0.21	0.22	65.71	0.95	0.777	-0.01	0.22	63.76	-0.03	1.000
M	p,t,k – dev.	0.18	0.15	71.45	1.19	0.636	1.09	0.15	79.56	6.91	<.0001*
	p,t,k – vcd.	0.66	0.15	68.42	4.44	0.0002*	1.20	0.14	58.91	8.42	<.0001*
	p,t,k – m	0.79	0.24	67.15	3.32	0.008*	1.22	0.24	65.06	5.15	<.0001*
	dev. – vcd.	0.48	0.11	115.86	4.34	0.0002*	0.11	0.11	103.22	1.02	0.738
	dev. – m	0.61	0.22	71.54	2.77	0.035	0.12	0.22	72.13	0.56	0.945
	vcd. – m	0.13	0.22	67.03	0.60	0.932	0.01	0.21	59.97	0.06	1.000

Table 18 ***Pairwise comparisons of z-score f0 after word-medial voiceless vs. voiced plosives (carrier sentence) (over the first 10% of the vowel) by tone context***

			WM_CARRIER				
	Tone	Contrast	*β*	SE	df	t.ratio	p.value
F	L	p,t,k – vcd.	-0.02	0.45	21.11	-0.04	0.999
	H	p,t,k – vcd.	0.46	0.56	18.42	0.82	0.697
M	L	p,t,k – vcd.	0.24	0.45	20.19	0.54	0.852
	H	p,t,k – vcd.	0.34	0.56	18.47	0.61	0.819

3.5 f0 perturbations: no compromise on pitch-accent pattern

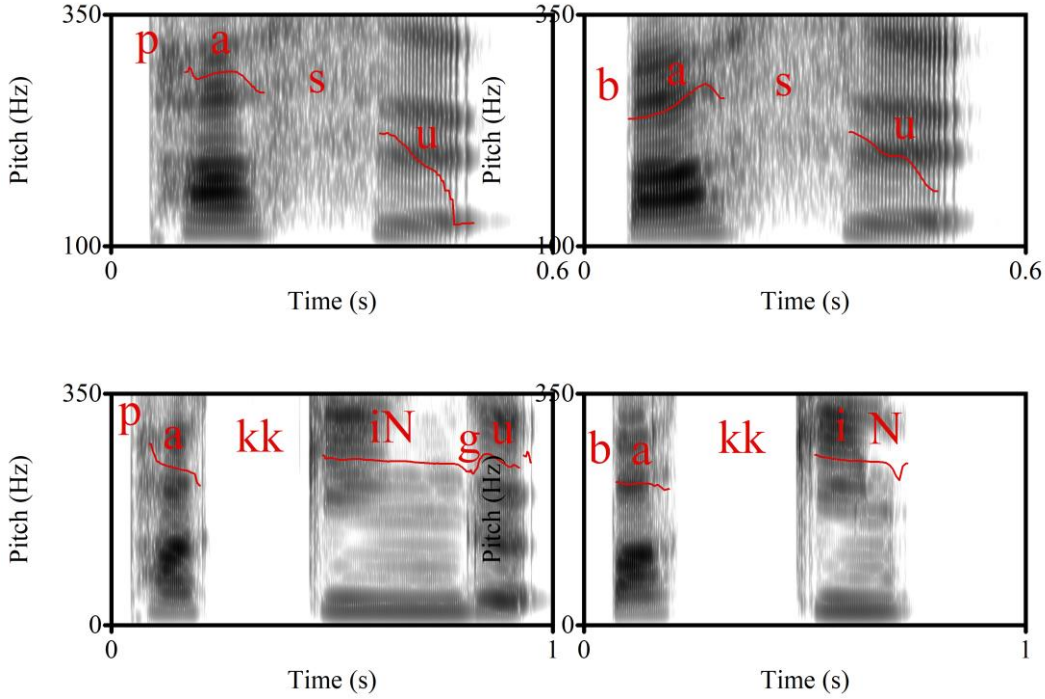


Figure 6 ***f0 curves of an HL minimal pair (upper panel) and an LH near-minimal pair (lower panel) with word-initial /p-b/ contrast in WI_citation***

Despite the *f*0 perturbation effect in word-initial position, the pitch-accent pattern was preserved notably by means of the offset *f*0 of the initial mora relative to the following *f*0 height. As shown in Figure 5abc, at the onset of the initial mora, the *f*0 height is at about the same level after voiced plosives in the H tone context as after voiceless plosives in the L tone context. However, at the endpoint of the mora, the *f*0 curves converge at a higher level in the H tone context, and at a lower level in the L tone context. As shown by the two (near-)minimal pairs in Figure 6, the initial mora /ba/ exhibits a much lower *f*0 than /pa/ at the vowel onset. However, at the vowel offset, the *f*0 difference is canceled. Importantly, the pitch-accent patterns, that is, HL in the upper panel, and LHHH(H) in the lower panel, remain unaffected.

3.6 Correlation between VOT and onset f0

Figure 7 plots the VOT of plosives against the *f*0 in semitone at the first time interval of the following vowel in WI_CITATION. Similar patterns can be found in WI_CARRIER-FOCUS. The regression lines fitted within each phonological voicing category, as well as the coefficients of Pearson's product-moment correlation test (Table 19), show a positive correlation between VOT and onset *f*0 within the voiced category in both the L and H tone contexts for both males and females. However, individual variations can be observed in both the male and female groups, as shown in the individual plots (Appendix 5). In addition, for prevoiced plosives ([b, d, g] in Table 19), no correlation was found between the length of the voice lead and onset *f*0. As for the voiceless category, female speakers show a negative correlation between VOT and onset *f*0 in L but not H tone contexts. Male speakers instead show a positive correlation between VOT and onset *f*0 within the voiceless category. Again, individual variations can be observed in both the male and female groups (Appendix 5). Therefore, we may only conclude that there is an inconsistency in either the presence or absence of this correlation.

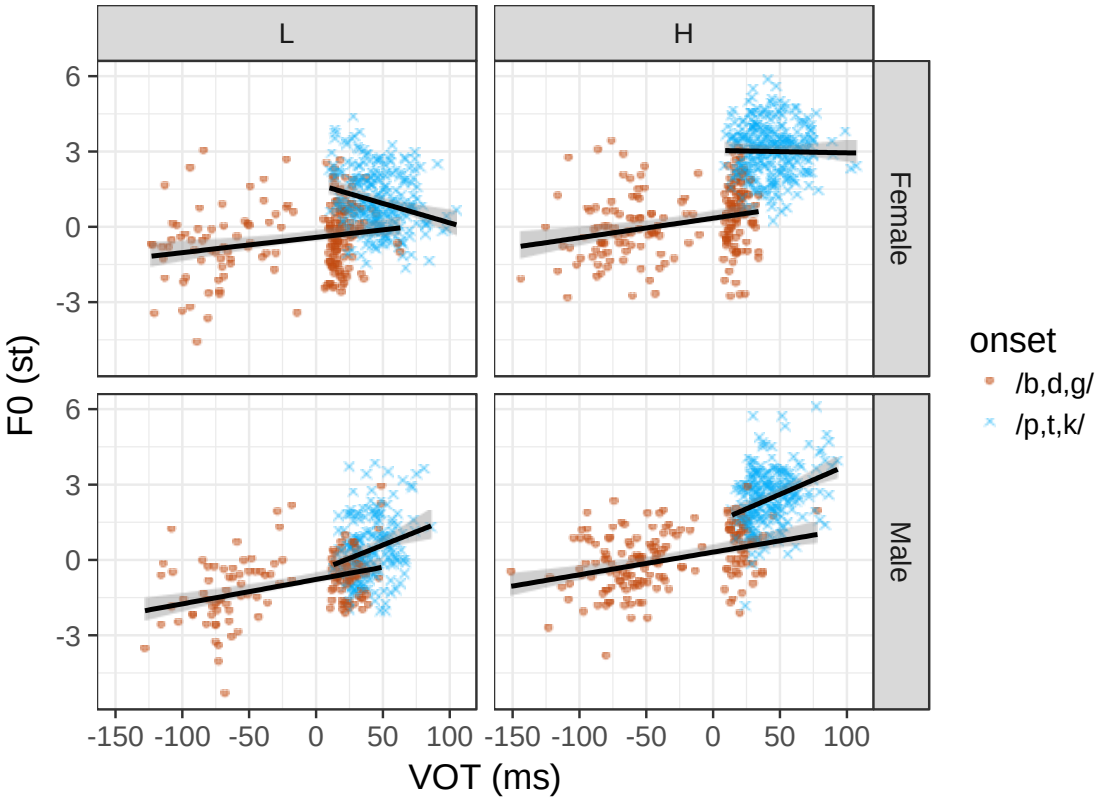


Figure 7 ***Scatter plots of VOT in ms against onset f0 in semitone in WI_citation, by tone and phonological voicing, with regression lines fitted within each phonological voicing category***

Table 19 ***Results of Pearson's product-moment correlation between VOT and onset f0 in semitone in WI_citation, by tone context and voicing category, * for*** **p** ***< .05***

	L			H		
	/p, t, k/	/b, d, g/	[b, d, g]	/p, t, k/	/b, d, g/	[b, d, g]
Female	*r* = -0.24, *p* < .0005 *	*r* = 0.22, *p* < .005 *	*r* = 0.24, *p* =.50	*r* = -0.02, *p* = .80	*r* = 0.27, *p* < .0005 *	*r* = -0.07, *p* = .78
Male	*r* = 0.25, *p* < .005 *	*r* = 0.40, *p* < .0001 *	*r* = -0.01, *p* = .96	*r* = 0.34, *p* < .0001 *	*r* = 0.36, *p* < .0001 *	*r* = 0.26, *p* = .96

4. Discussion

In this study, we explored acoustic cues (VOT, voicing-ratio, closure duration, *f*0) that contribute to the distinction between the two plosive series in Tokyo Japanese. The realizations of these cues differ essentially according to the position in the word. It is possible that word-initial and word-medial positions would be better described as Accentual Phrase-initial and Accentual Phrase-medial. However, since our analyses were based on lexical words only, we refrain from generalization to other units.

Voiced plosives are frequently devoiced in word-initial, and especially post-pausal position, resulting in overlapping VOTs with voiceless plosives. The realization of voiced plosives in intersonorant word-initial position meets the criterion of "passive voicing" based on Beckman et al. (2013). On the other hand, in word-initial position, voiceless series are frequently realized with moderate aspiration. It seems, however, that neither prevoicing nor aspiration allows for a *robust* distinction between the two plosive series word-initially. This distinctiveness is enhanced through *f*0 on the following vowel: *f*0 is higher after /p, t, k/, phonetically aspirated or unaspirated, than after /b, d, g/, phonetically voiced or voiceless.

In word-medial position, prevoicing is frequently present in the production of voiced plosives, whereas voiceless plosives are unaspirated. Voicing-ratio robustly distinguishes the two plosive series. The phonetic voicing of voiced plosives also results in a much shorter closure duration than voiceless plosives. No *f*0 difference is observed on the following vowel.

4.1 Between "true voicing" and "aspirating"

We examined the implementation of the laryngeal contrast in Tokyo Japanese. We will now relate our results to the expectations of a typical "true voicing" language, and to those of a typical "aspirating" language, listed in Section 1.3.

In POST-PAUSAL position (WI_CITATION and WI_CARRIER-FOCUS), phonologically voiced plosives are commonly devoiced and realized as voiceless unaspirated. The devoicing rate is remarkably higher than is usually observed in a "true voicing" language. This outcome fits well with (B1), expected for a typical "aspirating" language:

(B1) In post-pausal position, voiced plosives are not robustly produced with prevoicing, but are most likely to be voiceless unaspirated, as indicated by short-lag VOTs (mostly between 0 and 30 ms).

Phonologically voiced plosives are often prevoiced in INTERSONORANT position (WI_CARRIER, WM_CITATION, WM_CARRIER). However, the percentage of fully voiced tokens in WI_CARRIER in Tokyo Japanese is 49.9%, lowered than 62.5% in German, commonly categorized as an "aspirating" language, and 97.5% in Russian, commonly categorized as a "true voicing" language (see Section 3.1). This outcome fits well with (B2), expected for a typical "aspirating" language:

(B2) In intersonorant word-initial position, voicing is, at most, passively maintained during the closure of voiced plosives, that is, a low percentage of voiced plosives are fully voiced.

Phonologically voiceless plosives are moderately aspirated in word-initial position, indicated by medium-lag VOTs. In word-medial position, they are generally unaspirated, indicated by short-lag VOTs. The lack of aspiration in word-medial voiceless plosives in Tokyo Japanese is similar to that observed in English in an unstressed word-medial position. (According to our unquantified observation on several recorded tokens, voiceless plosives followed by a high vowel may have very long VOTs and sound aspirated in word-medial position. Also, the following high vowel is sometimes devoiced.) In any case, VOTs are lengthened in word-initial compared to word-medial position. The effect of the prosodic hierarchy is, however, not directly shown in our study. This outcome partially fits with (B4), expected for a typical "aspirating" language:

(B4) VOTs of voiceless plosives are lengthened in domain-initial position.

This outcome does not fit well with either (A3) or (B3):

(A3) Voiceless plosives are most likely unaspirated, as indicated by short-lag VOTs (mostly between 0 and 30 ms);

(B3) Voiceless plosives are most likely aspirated, as indicated by long-lag VOTs (mostly longer than 50 ms).

Altogether, Tokyo Japanese follows some but not all of the patterns B1-4, suggesting that it is clearly not a "true voicing" language but also not a typical "aspirating" language. It could be that the separation between these two categories is not deterministic, and many languages with various patterns may be seen as belonging to an intermediate category, such as modern Hebrew (Raphael et al., 1995; see Section 1.1.1). It could also be that other cues are needed to complement VOT in the definition of the phonetic specification. Similarly to Tokyo Japanese, Swiss German uses closure duration and *f*0 of the following vowel as important secondary cues (Ladd & Schmid, 2018). Other cues may also include voice quality, duration of the preceding vowel, etc. (see Cho et al., 2019, for a review).
What we may conclude from Tokyo Japanese data is that an active [voice] feature at the phonological level is not necessarily associated with a "true voicing" category simply based on the VOT pattern of a language. It is, however, reasonable to assume that voiced plosives were truly voiced at a previous stage. Recent studies show that elderly speakers of modern Japanese produce negative VOTs with word-initial voiced plosives (Takada, 2011; Takada et al., 2015). Our data show that voiced plosives produced by young speakers are robustly prevoiced only in word-medial position. Only time will tell whether Tokyo Japanese will remain stable in this intermediate category or eventually shift to an "aspirating" language.

4.2. f0 perturbations: automatic or enhanced?
In Section 1.3, we also asked whether *f*0 perturbations in Tokyo Japanese were an automatic or a controlled, enhanced effect, and whether the articulatory source was *f*0 raising or *f*0 lowering. The pattern of *f*0 perturbations is clearly dependent on the position in the word. In word-initial position, in which the sole VOT cue is not sufficient to distinguish the two plosive series, *f*0 perturbations are quite large in magnitude and duration. The duration of *f*0 perturbations is not limited to the vowel onset, but actually extends up to the final part of the vowel, suggesting that *f*0 perturbations are not inhibited by an existing "pitch-accent," or word-tone system, contrary to the patterns found in several tone languages (see Section 1.1.2). In word-medial position, the *f*0 perturbation effect is negligible, whereas voicing-ratio is sufficient to distinguish the two plosive series. These outcomes fit well with (Z2), expected for a controlled *f*0 perturbation account:
(Z2) *f*0 perturbations are larger in contexts in which the primary voicing cue is less reliable.

The following are the observations regarding the word-initial position only.
*f*0 is higher after voiceless plosives than after prevoiced plosives, devoiced plosives, and /m/. This outcome fits well with (Y1), expected for an automatic *f*0 raising account:
(Y1) *f*0 is raised after voiceless plosives compared to the /m/ onset.

*f*0 perturbations show the same pattern after phonologically voiced plosives, whether prevoiced or devoiced, comparable to the results in English and Spanish (Dmitrieva et al., 2015). This outcome contradicts (X2), expected for an automatic *f*0 lowering account:
(X2) When voiced plosives are produced without closure voicing, *f*0 is raised; in other words, devoiced plosives are followed by higher *f*0 than prevoiced plosives.
It also fits well with (Z1), expected for a controlled *f*0 perturbation account:
(Z1) *f*0 perturbations are conditioned by the phonological voicing contrast, but not predicted by VOT: a similar *f*0 perturbation effect is observed regardless of the closure voicing of voiced plosives.

A positive correlation between onset *f*0 and VOT is found within the voiced category, with individual variations. Within the voiceless category, a negative correlation between *f*0 and VOT is found only in the L tone context for female speakers overall, again with high variability. No clear pattern can be found for (X4), expected for an automatic *f*0 lowering account, nor for (Z3), expected for a controlled *f*0 account:

(X4) Onset *f*0 correlates positively with VOT in the negative VOT range: the longer the prevoicing (i.e., smaller VOT), the lower the onset *f*0 value;
(Z3) Within each phonological voicing category, onset *f*0 correlates negatively with VOT: the smaller the VOT value, the higher the onset *f*0 value.

Lastly, *f*0 perturbations are much larger in high than low tone contexts. The effect of *f*0 context has also been found in other languages, such as German (Kohler, 1982), English (Hanson, 2009), French and Italian (Kirby & Ladd, 2016). Hanson (2009) interpreted this difference as a conflict between the prosodic gesture for lowering *f*0 and the segmental gesture for inhibiting voicing. Indeed, the laryngeal muscle activities associated with *f*0 production are not the same in different *f*0 ranges. Electromyographic (EMG) data on Thai and Chinese tones (Erickson, 1993; Hallé, 1994) demonstrated that the cricothyroid muscle (CT) is activated in high or raised *f*0 (but not in the low *f*0 range) while strap muscles are activated to lower *f*0 (and variably in the mid *f*0 range, but not in the high *f*0 range). If the *f*0 perturbation effect in Tokyo Japanese is indeed caused by *f*0 raising due to CT stiffening, which is more favored in higher than lower *f*0 ranges, this might explain why *f*0 is more perturbed in high than low *f*0 contexts. It is interesting to note that similar effects have been found with the intrinsic *f*0 related to vowel height, that is, the *f*0 difference between high and low vowels is larger in higher than in lower *f*0 ranges (Whalen & Levitt, 1995). However, the physiological mechanisms might not be the same: the authors attributed this variation to the conflicting gestures of strap muscles and the tongue pulling that affects the hyoid bone in low *f*0 ranges. What is clear is that laryngeal mechanisms involving segmental and *f*0 productions are complex and interdependent. Further investigations are needed to explore synergistic and conflicting laryngeal gestures at the segmental and suprasegmental levels.
Overall, our data lend more support to an effect of *f*0 raising of voiceless plosives as the articulatory source of *f*0 perturbations. Without being able to further disentangle what is automatic from what is controlled, we adopt the hybrid model (Hoole & Honda, 2011; Dmitrieva et al., 2015) to explain the *f*0 perturbation effect in Tokyo Japanese: there is an automatic component in the *f*0 perturbation effect, mostly likely due to raised *f*0 after voiceless plosives, but *f*0 can also be enhanced, that is, controlled deliberately, in some contexts (also see Section 4.3).

4.3. f0 enhancements and implications for sound change
Our young speakers' data corroborate previous studies' findings (Takada, 2011; Takada et al., 2015), in that these speakers very often omit closure voicing during the production of voiced plosives in word-initial position. In this position, the moderate aspiration produced for voiceless plosives compared to the lack of aspiration for voiced plosives cannot be viewed as a large difference. Indeed, our read speech data show an overlap between the two plosive series in terms of VOT. This overlap would probably be even greater in connected or hypoarticulated speech. Because the reliability of the primary VOT cue to voicing therefore seems weak, speakers are more likely to use secondary cues to maintain the voicing contrast.
The qualitative difference of *f*0 perturbations and voicing pattern between word-initial and word-medial positions suggests that *f*0 perturbations are enhanced when the primary cue is weakened, that is, less reliable because of possible confusion based on this sole cue. Kingston and Diehl (1994) attributed *f*0 lowering to a deliberate reinforcement of low-frequency spectral energy to help identify the [+voice] feature. However, we believe that Tokyo Japanese speakers are not enhancing the percept of [+voice] feature only, but are recovering the blurred out contrast by using other strategies, including aspiration of voiceless plosives and *f*0 difference on the following vowel. In other words, secondary cues are enhanced to *recover* the *contrast* in cases where the primary cue is not sufficiently distinctive due to such factors as hypoarticulation, articulatory challenges (e.g., Aerodynamic Voicing Constraints, Ohala, 2011), and perceptual confusion. This view of enhancement is closer to Silverman's (1997), although he placed greater emphasis on contrast recovery due to morpho-phonological constraints. Importantly, the *f*0 cue may replace the VOT cue, leading to transphonologization (Hagège & Haudricourt, 1978; Hyman, 2013), which is the basis of tone split (Haudricourt, 1961).

Our study does not provide direct evidence for a shift in cue weighting — in production or in perception — in Tokyo Japanese, whether it be an increased distinctiveness in the *f*0 difference, or an increased aspiration in voiceless plosives in parallel with a decreased prevoicing in the voiced plosives, or both. That said, the average magnitude of *f*0 perturbations in hertz after word-initial plosives is comparable between our data and Shimizu's (1996) data, which might suggest that the magnitude has not evolved much during the last 20-30 years. Our Tokyo Japanese patterns are quite similar to Afrikaans (Coetzee et al., 2018, Section 1.1.2): (a) of the two plosive series, voiced plosives are devoiced in word-initial position in young speakers' production; (b) the *f*0 perturbation effect has a fairly large magnitude, extending through the entire vowel, but no increase in the *f*0 difference is observed across generations. The authors did not firmly conclude, but still suggested, that Afrikaans might be in the course of an ongoing sound change of voicing contrast replaced by a tonal contrast.

In the same vein, we might also consider whether a tonal development could take place in Tokyo Japanese. If Tokyo Japanese represents the very initial stage of a tonal development, or even prior to this stage, the reduced VOT contrast in word-initial position might be a precursor of such a sound change, triggering the enhancement of *f*0. Similarly, it is shown that the origin of the incipient tonogenesis in Seoul Korean (not its propagation) lies in a *production bias* to reduce VOT contrast due to hypoarticulation (Bang, Sonderegger, Kang, Clayards, & Yoon, 2018). *f*0 is a very useful "spare wheel" because of its perceptual salience and the multiple parameters (height, contour, turning point, timing, etc.) that can be manipulated. As Hyman (2011) put it, "Tone can do everything that segmental and metrical phonology can do, but the reverse is not true." Domain-initial devoicing is frequently observed in many languages, including Japanese, English, and Afrikaans. However, not all languages will develop tones. Our speculation is that the possibility to develop tones does not solely rely on the cue weighting relation between *f*0 and another cue such as VOT; other conditions in the phonological system may either propel the gradual replacement of VOT by *f*0, or put a brake on this change. We further speculate that such a sound change is slowed down in Tokyo Japanese by two factors: (a) instead of a three-way plosive contrast, as in Seoul Korean, Japanese has a two-way plosive contrast, leaving more space for a shift solely in the VOT dimension, thus, aspiration also contributes to recovering the endangered contrast, as our results showed; and (b) *f*0 perturbations are restrained from extending beyond the initial mora so that the relative *f*0 height with the following mora is maintained for the preservation of the pitch-accent pattern. Testing of these speculations must be left to future work; yet, we believe that a model of sound change should consider not only how the target sounds change but also how they resist change due to the constraints of the entire phonological system.

4.4. Enhancement of prevoicing and aspiration

As a final note on enhancement, another observation can be made on the basis of our data. In word-initial position, the rate of prevoicing is higher overall in H than L tone contexts (see Section 3.1), while aspiration is longer in L than H tone contexts, but for female speakers only (see Section 3.2). Martine Grice (personal communication, 2018) mentioned that devoicing is more common in low *f*0 environments, as shown by a much higher vowel devoicing rate in low than high tone contexts in Japanese (also see our results in Section 2.4). However, we are not aware of any physiological mechanisms that would motivate prevoicing in the high *f*0 context. Louis Goldstein (personal communication, 2018) mentioned that if larynx lowering and the adduction of vocal folds are both involved in voicing, speakers might tend to use more adduction gestures to compensate for the absence of larynx lowering in the high *f*0 range. Likewise, we might also reason that in the production of voiceless plosives in the low *f*0 range, the stiffening gesture is compromised, and thus speakers might tend to enhance the spread gesture for compensation, which might result in a longer VOT. Similar enhancements have been suggested in Hanson (2009), although she predicted a higher probability of closure voicing for voiced plosives in low than high *f*0 contexts.

An alternative explanation may be considered. So far, we have been focusing on the distinctiveness between the two plosive series as well as the distinctiveness between initial H and L tones. If we list the acoustic cues of voiceless vs. voiced plosives in H vs. L tone contexts

in word-initial position (Table 20), we observe that there is a possible confusion between voiceless plosives in the L tone context and voiced plosives in the H tone context, because their tone and *f*0 perturbation effect tend to cancel each other, at least at the vowel onset. It is thus possible that (some) speakers try to avoid this confusion by enhancing prevoicing to signal voicedness in the H tone context, and longer aspiration to signal voicelessness in the L tone context. This speculation needs further investigation as well as perceptual evidence to confirm it. If it is true, it further supports the idea that multiple cues, rather than one active feature or one property, can be enhanced for the purpose of recovering a contrast. Speakers might simply use the handiest cues, that is, cues that are both articulatorily effortless and perceptually salient. These cues may differ from one context to another and may be subject to physiological constraints and individual variations.

Table 20 ***Multiple cues correlated to word-initial voiceless vs. voiced plosives in H vs. L tone. "++" stands for more frequent prevoicing, or longer aspiration compared to the other pitch-accent context***

	/p, t, k/	/b, d, g/
H	H tone higher *f*0 moderate aspiration	
		H tone **lower *f*0** **variable prevoicing++**
L	**L tone** **higher *f*0** **moderate aspiration++**	L tone lower *f*0 variable prevoicing

5. Conclusions

The main goal of this study was to examine voicing-related and *f*0 properties associated with the laryngeal contrast of plosives in modern Tokyo Japanese. Our VOT and voicing-ratio data showed that (a) voiced plosives exhibit a high devoicing rate in post-pausal position, a passive voicing pattern in intersonorant word-initial position, and robust prevoicing in word-medial position; and (b) voiceless plosives are variably and moderately aspirated in word-initial position and unaspirated in word-medial position. Contrary to the (morpho-)phonological proposal of an active [voice] feature in Japanese, our phonetic data suggest that Japanese is not a "true voicing" language, and that VOT alone is insufficient in distinguishing its two plosive series. Frequent devoicing of word-initial voiced plosives leads to an overlap in VOT distribution with voiceless plosives. On the other hand, in word-initial position, *f*0 of the following vowel is higher after phonologically voiceless than phonologically voiced obstruents as well as nasals. This difference extends up to the final part of the vowel and is larger in H than L tone contexts. Conversely, in word-medial position, the *f*0 perturbation effect is negligible whereas both voicing-ratio and closure duration participate in the distinction of the two plosive series. This positional variation suggests that *f*0 perturbations in Tokyo Japanese are enhanced only in word-initial position, in which the primary voicing cue is not sufficient. The synchronic variations of voicing and *f*0 in Tokyo Japanese may provide insights into crosslinguistic implementations of the laryngeal contrast as well as possible sources of a potential tonal development.

Acknowledgements

This study was supported by a JSPS Postdoctoral Fellowship awarded to the first author, and was funded by JSPS Grant-in-Aid No. 17F17006. We wish to thank the editor, Taehong Cho, and three anonymous reviewers for their invaluable suggestions, which greatly improved this paper. The interpretation of the results benefited from numerous discussions with Martine Mazaudon. We are grateful to her as well as Marc Brunelle, Pierre Hallé, James Kirby, Mafuyu Kitahara, and Douglas Whalen, for very constructive comments on earlier drafts; to Boyd Michailovsky for proofreading the manuscript; and to Takeki Kamiyama for suggestions on some details of this study. None of them necessarily agrees on all the points reviewed or

developed in this paper. All errors are exclusively ours. Preliminary results were presented at the Spring Meeting of the Acoustical Society of Japan 2018, LabPhon 16, the 6th Symposium on Tonal Aspects of Languages, and the 16th Meeting of the French Phonology Network. Finally, thanks to all our participants, who were able to maintain admirably constant speech rate and intonation, making the analyses a lot easier!

Appendix 1. Wordlist

S1 is the target syllable.

	p	b	t	d	k	g	m
L	paːto	baːteN	taikeː	daikeː	kaikai	gaikai	maiko
	peːʣi	beːʣju	teki	deki	keːɕja	geːɕja	–
	–	–	–	–	kiseː	giseː	–
	pakkiŋgu	bakkiN	takkjuː	dakkjuː	kakki	gakki	–
	–	–	teppaN	deppaN	kekkaN	gekkaN	–
	peŋki	beŋkjoː	tentoː	dentoː	keŋkoː	geŋkoː	–
H	pasu	basu	taiɕi	daiɕi	kaimu	gaimu	maigo
	pesuto	besuto	teguʧi	deguʧi	keːsu	geːmu	–
	piru	biru	–	–	kiɕi	giɕi	–
	pakku	bakku	tattɕi	dattɕi	katto	gattsu	–
	–	–	tekkjo	dekki	kekkoː	gekka	–
	pentɕi	bentɕi	teŋki	deŋki	kenri	genri	–

S2 is the target syllable.

	p	b	t	d	k	g	m
L	kjampasu	kjambasu	setai ɕjotai	sedai[a] ɕjodai	sokai	sogai	samaː
H	kampai	kambai	ʣitai	ʣidai	ɕjoka	ɕjoga	seːmai

[a]This word was not read with the prescribed pitch-accent by most female speakers, so the word pair [setai] – [sedai] was replaced with /ɕjotai/ – /ɕjodai/ in later recordings with male speakers.

Appendix 2. Summaries of statistical models

MODEL 1: voiced ~ (position + POA | subject) + (1 | item) + sex + position + pitch.accent + POA + vowel + syllable + sex:position + sex:pitch.accent + position:pitch.accent + sex:POA + sex:vowel + sex:syllable + sex:position:pitch.accent

Family: binomial (logit)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.81634	0.55937	-3.247	0.001166	**
sexM	1.04334	0.78084	1.336	0.181488	
positionWIcarrier(focus)	1.18137	0.42866	2.756	0.005852	**
positionWIcarrier	2.45774	0.78092	3.147	0.001648	**
positionWMcarrier	4.76784	0.94325	5.055	4.31e-07	***
positionWMcitation	5.10706	1.09732	4.654	3.25e-06	***
pitch.accentH	0.28995	0.22269	1.302	0.192903	
POAd	0.02901	0.32350	0.090	0.928533	
POAg	0.05685	0.39213	0.145	0.884735	
vowele	1.19638	0.20903	5.723	1.04e-08	***
voweli	1.65499	0.32650	5.069	4.00e-07	***
syllableCVQ	-0.19158	0.20657	-0.927	0.353690	
syllableCVN	-0.11769	0.24593	-0.479	0.632248	
sexM:positionWIcarrier(focus)	-0.67642	0.60964	-1.110	0.267197	
sexM:positionWIcarrier	-0.67348	1.01759	-0.662	0.508074	
sexM:positionWMcarrier	-0.39203	1.32106	-0.297	0.766656	
sexM:positionWMcitation	-2.12169	1.40033	-1.515	0.129738	
sexM:pitch.accentH	0.79071	0.30856	2.563	0.010388	*
positionWIcarrier(focus):pitch.accentH	0.48150	0.35054	1.374	0.169567	
positionWIcarrier:pitch.accentH	0.03515	0.45112	0.078	0.937902	
positionWMcarrier:pitch.accentH	-0.23574	0.85150	-0.277	0.781896	
positionWMcitation:pitch.accentH	-0.60049	0.82811	-0.725	0.468372	
sexM:POAd	0.07406	0.44436	0.167	0.867636	
sexM:POAg	-0.07763	0.54463	-0.143	0.886663	
sexM:vowele	-0.77567	0.27640	-2.806	0.005011	**
sexM:voweli	-1.52145	0.42825	-3.553	0.000381	***
sexM:syllableCVQ	0.06808	0.27409	0.248	0.803837	
sexM:syllableCVN	0.25399	0.33422	0.760	0.447280	
sexM:positionWIcarrier(focus):pitch.accentH	-0.99923	0.54351	-1.838	0.065994	.
sexM:positionWIcarrier:pitch.accentH	0.12503	0.60993	0.205	0.837579	
sexM:positionWMcarrier:pitch.accentH	-2.25689	1.20825	-1.868	0.061776	.
sexM:positionWMcitation:pitch.accentH	-0.26634	1.04052	-0.256	0.797976	

MODEL 2: VOT ~ (position | subject) + (1 | item) + sex + position + pitch.accent + POA + vowel + sex:position + sex:pitch.accent + position:pitch.accent + sex:POA + sex:vowel + sex:position:pitch.accent

	Estimate	Std. Error	t value
(Intercept)	63.4276	3.3026	19.205
sexM	-5.5374	3.9180	-1.413
positionWIcarrier(focus)	-5.0246	2.3147	-2.171
positionWIcarrier	-6.3810	2.1897	-2.914
positionWMcarrier	-34.8517	3.7423	-9.313

positionWMcitation	-34.7188	3.8626	-8.988
pitch.accentH	-6.0880	1.7987	-3.385
POAp	-18.3780	1.9097	-9.623
POAt	-20.1356	1.8692	-10.772
vowele	-7.0925	1.7981	-3.944
voweli	10.6756	3.2511	3.284
sexM:positionWIcarrier(focus)	4.1523	3.4203	1.214
sexM:positionWIcarrier	6.3201	2.9713	2.127
sexM:positionWMcarrier	10.5823	3.6670	2.886
sexM:positionWMcitation	10.2924	3.9573	2.601
sexM:pitch.accentH	3.2216	1.1462	2.811
positionWIcarrier(focus):pitch.accentH	0.8723	1.3300	0.656
positionWIcarrier:pitch.accentH	-0.8225	1.4927	-0.551
positionWMcarrier:pitch.accentH	8.9795	4.2848	2.096
positionWMcitation:pitch.accentH	8.5318	4.2499	2.008
sexM:POAp	-2.3713	0.9154	-2.590
sexM:POAt	1.7897	0.9182	1.949
sexM:vowele	3.3128	0.8561	3.870
sexM:voweli	2.2767	1.7633	1.291
sexM:positionWIcarrier(focus):pitch.accentH	0.5281	2.0583	0.257
sexM:positionWIcarrier:pitch.accentH	2.7789	2.0284	1.370
sexM:positionWMcarrier:pitch.accentH	-4.6227	2.9202	-1.583
sexM:positionWMcitation:pitch.accentH	-4.9897	2.8843	-1.730

MODEL 3: VOT ~ (position + POA + POA + pitch.accent | subject) + (1 | item) + POA + position + POA + sex + vowel + syllable + pitch.accent + POA:position + POA:POA + position:POA + POA:sex + position:sex + POA:sex + sex:vowel + sex:syllable + sex:pitch.accent + POA:position:POA + POA:position:sex + POA:POA:sex + position:POA:sex + POA:position:POA:sex

	Estimate	Std. Error	t value
(Intercept)	28.68637	2.14459	13.376
VOICEvoiceless	30.19783	2.99300	10.089
positionWIcitation	2.69800	2.00021	1.349
POAp	-8.07979	2.69404	-2.999
POAt	-6.39817	2.64237	-2.421
sexM	5.78685	3.04891	1.898
vowele	-4.40163	1.10703	-3.976
voweli	4.62733	1.84422	2.509
syllableCVQ	-5.41318	1.06914	-5.063
syllableCVN	-3.05688	1.38019	-2.215
pitch.accentH	-3.88258	1.20729	-3.216
VOICEvoiceless:positionWIcitation	4.27361	1.82383	2.343
VOICEvoiceless:POAp	-9.45286	2.97085	-3.182
VOICEvoiceless:POAt	-12.44766	2.93215	-4.245
positionWIcitation:POAp	-3.74821	2.28369	-1.641
positionWIcitation:POAt	-3.47616	2.27479	-1.528
VOICEvoiceless:sexM	-1.34353	4.24856	-0.316

positionWIcitation:sexM	-2.44557	3.31420	-0.738
POAp:sexM	-5.38755	3.87153	-1.392
POAt:sexM	-7.50976	4.08695	-1.837
sexM:vowele	2.19172	0.87277	2.511
sexM:voweli	1.46042	1.54751	0.944
sexM:syllableCVQ	-0.86273	0.87369	-0.987
sexM:syllableCVN	1.26225	1.09820	1.149
sexM:pitch.accentH	2.02076	1.32931	1.520
VOICEvoiceless:positionWIcitation:POAp	-1.01725	2.70252	-0.376
VOICEvoiceless:positionWIcitation:POAt	-0.93155	2.66556	-0.349
VOICEvoiceless:positionWIcitation:sexM	-6.57118	3.18268	-2.065
VOICEvoiceless:POAp:sexM	2.24070	3.95117	0.567
VOICEvoiceless:POAt:sexM	5.10238	4.16235	1.226
positionWIcitation:POAp:sexM	1.13701	3.92616	0.290
positionWIcitation:POAt:sexM	5.21496	4.16292	1.253
VOICEvoiceless:positionWIcitation:POAp:sexM	-0.20819	4.51967	-0.046
VOICEvoiceless:positionWIcitation:POAt:sexM	-0.08859	4.66687	-0.019

MODEL 4: C_vratio ~ (pitch.accent | subject) + (1 | item) + POA + sex + position + POA + POA:sex + POA:position + sex:position + sex:POA + position:POA + POA:sex:position + sex:position:POA

	Estimate	Std. Error	t value
POAp	0.0851922	0.0270951	3.144
POAt	0.0913872	0.0264370	3.457
sexM	0.0572025	0.0334075	1.712
positionWMcitation	0.2232465	0.0484773	4.605
positionWMcarrier	0.2335489	0.0485781	4.808
VOICEvoiceless	-0.5782732	0.0221501	-26.107
POAp:sexM	-0.0233207	0.0230709	-1.011
POAt:sexM	-0.0647982	0.0226853	-2.856
POAp:positionWMcitation	-0.0781089	0.0599217	-1.304
POAt:positionWMcitation	-0.0317083	0.0582099	-0.545
POAp:positionWMcarrier	-0.0611645	0.0599217	-1.021
POAt:positionWMcarrier	-0.0010588	0.0592331	-0.018
sexM:positionWMcitation	-0.1379577	0.0353885	-3.898
sexM:positionWMcarrier	-0.0141398	0.0354609	-0.399
sexM:VOICEvoiceless	0.0312033	0.0189700	1.645
positionWMcitation:VOICEvoiceless	-0.2482199	0.0473022	-5.248
positionWMcarrier:VOICEvoiceless	-0.2474358	0.0476530	-5.192
POAp:sexM:positionWMcitation	-0.0005984	0.0426417	-0.014
POAt:sexM:positionWMcitation	0.0479985	0.0455514	1.054
POAp:sexM:positionWMcarrier	-0.0575127	0.0424746	-1.354
POAt:sexM:positionWMcarrier	-0.0757689	0.0466988	-1.623
sexM:positionWMcitation:VOICEvoiceless	0.0841983	0.0361697	2.328
sexM:positionWMcarrier:VOICEvoiceless	0.0562845	0.0363958	1.546

MODEL 5: closure.dur ~ (position + pitch.accent | subject) + (1 | item) + POA + sex + position + POA + vowel + pitch.accent + POA:sex + POA:position + sex:position + sex:POA + position:POA + sex:pitch.accent + POA:sex:position + sex:position:POA

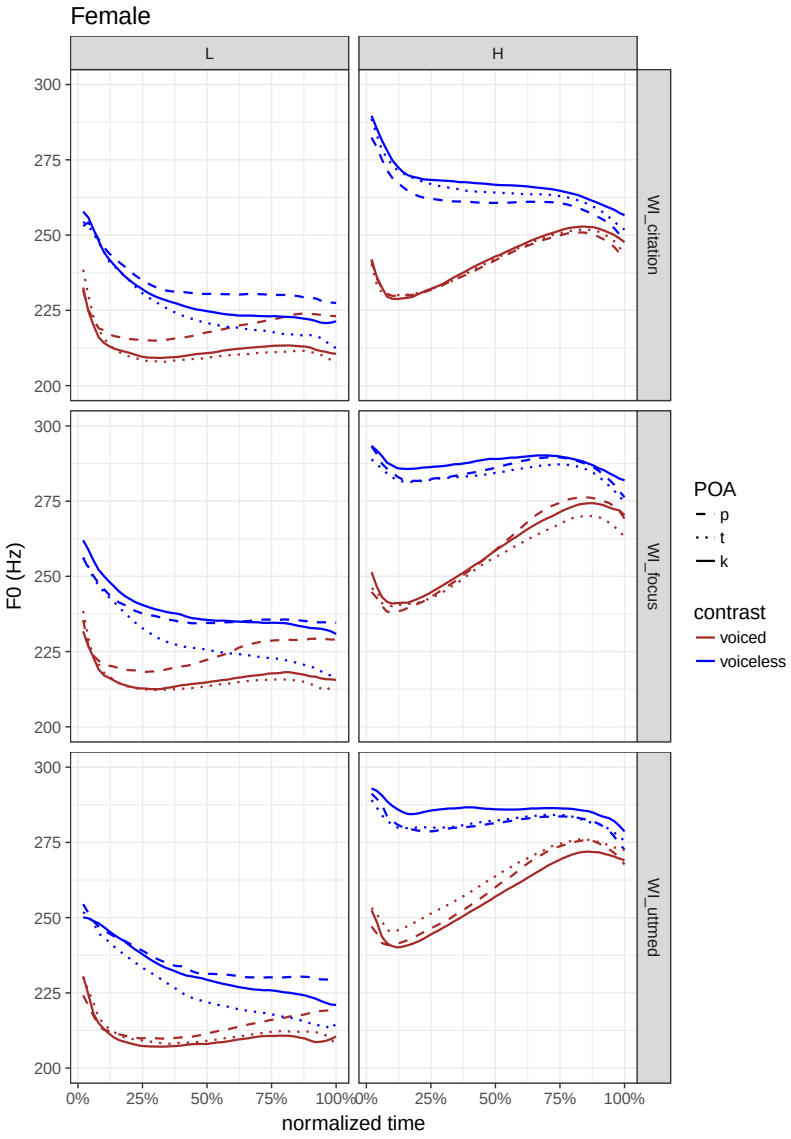
	Estimate	Std. Error	t value
(Intercept)	101.8027	16.8610	6.038
POAp	7.1971	2.1164	3.401
POAt	-3.6492	2.0935	-1.743
sexM	-22.2982	22.3294	-0.999
positionWMcitation	-59.3420	16.4007	-3.618
positionWMcarrier	-72.0771	16.6361	-4.333
VOICEvoiceless	19.1655	1.7385	11.024
vowele	1.9142	1.3861	1.381
voweli	7.3393	2.6052	2.817
pitch.accentH	-0.8106	2.2730	-0.357
POAp:sexM	0.1601	2.4729	0.065
POAt:sexM	-0.6229	2.4144	-0.258
POAp:positionWMcitation	-18.7045	4.4690	-4.185
POAt:positionWMcitation	-5.9550	4.6271	-1.287
POAp:positionWMcarrier	-11.6208	4.6026	-2.525
POAt:positionWMcarrier	4.3366	4.8062	0.902
sexM:positionWMcitation	16.4559	21.4917	0.766
sexM:positionWMcarrier	20.7871	21.8686	0.951
sexM:VOICEvoiceless	-0.5411	2.0277	-0.267
positionWMcitation:VOICEvoiceless	48.6246	3.8616	12.592
positionWMcarrier:VOICEvoiceless	29.3408	4.0446	7.254
sexM:pitch.accentH	-1.4439	2.8551	-0.506
POAp:sexM:positionWMcitation	5.2350	5.0585	1.035
POAt:sexM:positionWMcitation	5.6294	5.2948	1.063
POAp:sexM:positionWMcarrier	7.4197	5.2543	1.412
POAt:sexM:positionWMcarrier	3.7577	5.5275	0.680
sexM:positionWMcitation:VOICEvoiceless	-13.7874	4.4375	-3.107
sexM:positionWMcarrier:VOICEvoiceless	-15.4895	4.7227	-3.280

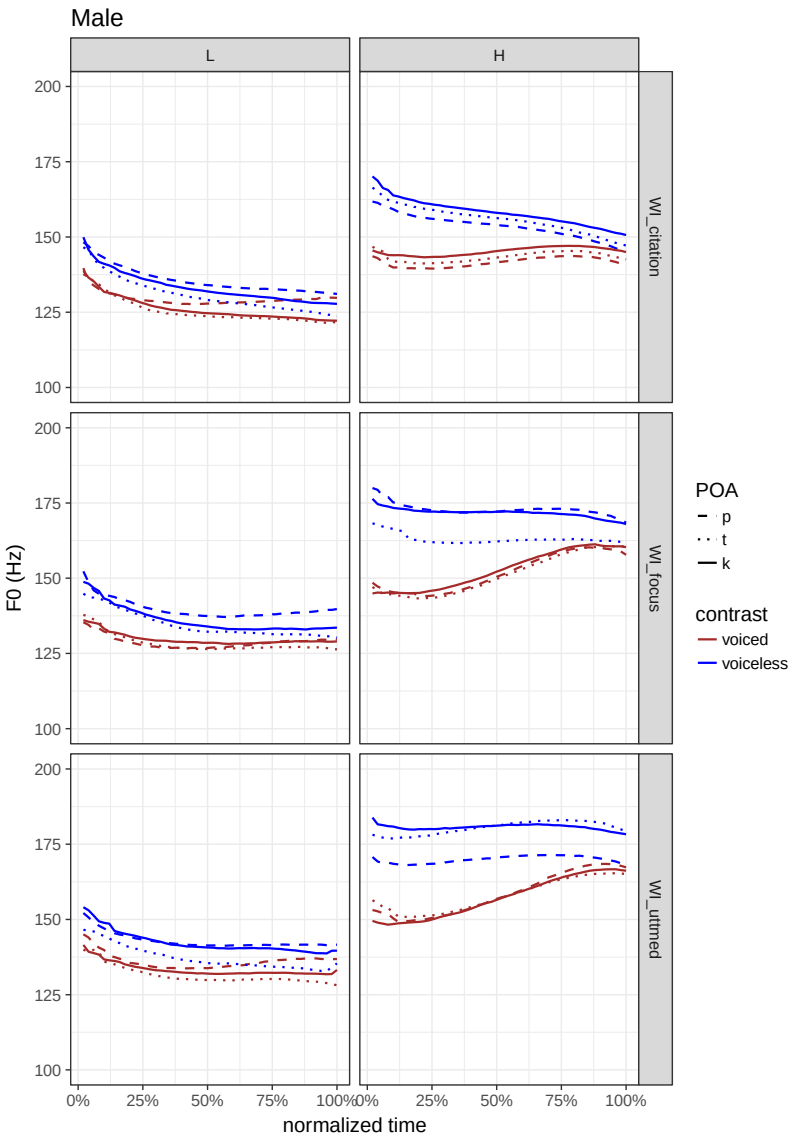
MODEL 6: f0.z ~ (onset.type + position | subject) + (1 | item) + sex + onset.type + position + pitch.accent + tpoint + sex:onset.type + sex:position + onset.type:position + sex:pitch.accent + onset.type:pitch.accent + position:pitch.accent + sex:onset.type:position + sex:onset.type:pitch.accent + sex:position:pitch.accent + onset.type:position:pitch.accent + sex:onset.type:position:pitch.accent

	Estimate	Std. Error	t value
(Intercept)	0.649977	0.121943	5.330
sexM	-0.313418	0.150397	-2.084
onset.typedevoiced	-1.067853	0.129848	-8.224
onset.typeprevoiced	-0.887953	0.136225	-6.518
onset.typem	-1.396428	0.207570	-6.727
positioncarrier-focus	-0.125103	0.097700	-1.280
positioncarrier	0.214265	0.103522	2.070
pitch.accentH	1.133478	0.088395	12.823
tpoint	-0.078143	0.003902	-20.027

sexM:onset.typedevoiced	0.371382	0.144351	2.573
sexM:onset.typeprevoiced	0.250767	0.148305	1.691
sexM:onset.typem	0.536025	0.193856	2.765
sexM:positioncarrier-focus	0.560236	0.145011	3.863
sexM:positioncarrier	0.041702	0.138550	0.301
onset.typedevoiced:positioncarrier-focus	0.254348	0.058907	4.318
onset.typeprevoiced:positioncarrier-focus	-0.037244	0.080709	-0.461
onset.typem:positioncarrier-focus	0.616728	0.090524	6.813
onset.typedevoiced:positioncarrier	-0.340786	0.092512	-3.684
onset.typeprevoiced:positioncarrier	-0.202089	0.091260	-2.214
onset.typem:positioncarrier	0.100702	0.106245	0.948
sexM:pitch.accentH	-0.058993	0.043368	-1.360
onset.typedevoiced:pitch.accentH	-0.719430	0.126853	-5.671
onset.typeprevoiced:pitch.accentH	-1.005704	0.140924	-7.137
onset.typem:pitch.accentH	-0.286124	0.233317	-1.226
positioncarrier-focus:pitch.accentH	0.342126	0.048419	7.066
positioncarrier:pitch.accentH	0.338001	0.053296	6.342
sexM:onset.typedevoiced:positioncarrier-focus	-0.270683	0.120105	-2.254
sexM:onset.typeprevoiced:positioncarrier-focus	-0.209074	0.116649	-1.792
sexM:onset.typem:positioncarrier-focus	-0.573094	0.139598	-4.105
sexM:onset.typedevoiced:positioncarrier	0.853595	0.124865	6.836
sexM:onset.typeprevoiced:positioncarrier	0.178348	0.123929	1.439
sexM:onset.typem:positioncarrier	-0.031585	0.149780	-0.211
sexM:onset.typedevoiced:pitch.accentH	0.360898	0.078060	4.623
sexM:onset.typeprevoiced:pitch.accentH	0.486479	0.098757	4.926
sexM:onset.typem:pitch.accentH	-0.163830	0.108006	-1.517
sexM:positioncarrier-focus:pitch.accentH	0.306437	0.078637	3.897
sexM:positioncarrier:pitch.accentH	0.209091	0.075085	2.785
onset.typedevoiced:positioncarrier-focus:pitch.accentH	-0.074258	0.083540	-0.889
onset.typeprevoiced:positioncarrier-focus:pitch.accentH	0.384372	0.105147	3.656
onset.typem:positioncarrier-focus:pitch.accentH	-0.870279	0.122117	-7.127
onset.typedevoiced:positioncarrier:pitch.accentH	0.416674	0.122288	3.407
onset.typeprevoiced:positioncarrier:pitch.accentH	0.622867	0.106173	5.867
onset.typem:positioncarrier:pitch.accentH	0.115614	0.140625	0.822
sexM:onset.typedevoiced:positioncarrier-focus:pitch.accentH	-0.497692	0.167066	-2.979
sexM:onset.typeprevoiced:positioncarrier-focus:pitch.accentH	-0.692326	0.150875	-4.589
sexM:onset.typem:positioncarrier-focus:pitch.accentH	0.399200	0.184742	2.161
sexM:onset.typedevoiced:positioncarrier:pitch.accentH	-0.969419	0.166630	-5.818
sexM:onset.typeprevoiced:positioncarrier:pitch.accentH	-0.647030	0.144387	-4.481
sexM:onset.typem:positioncarrier:pitch.accentH	-0.091162	0.196973	-0.463

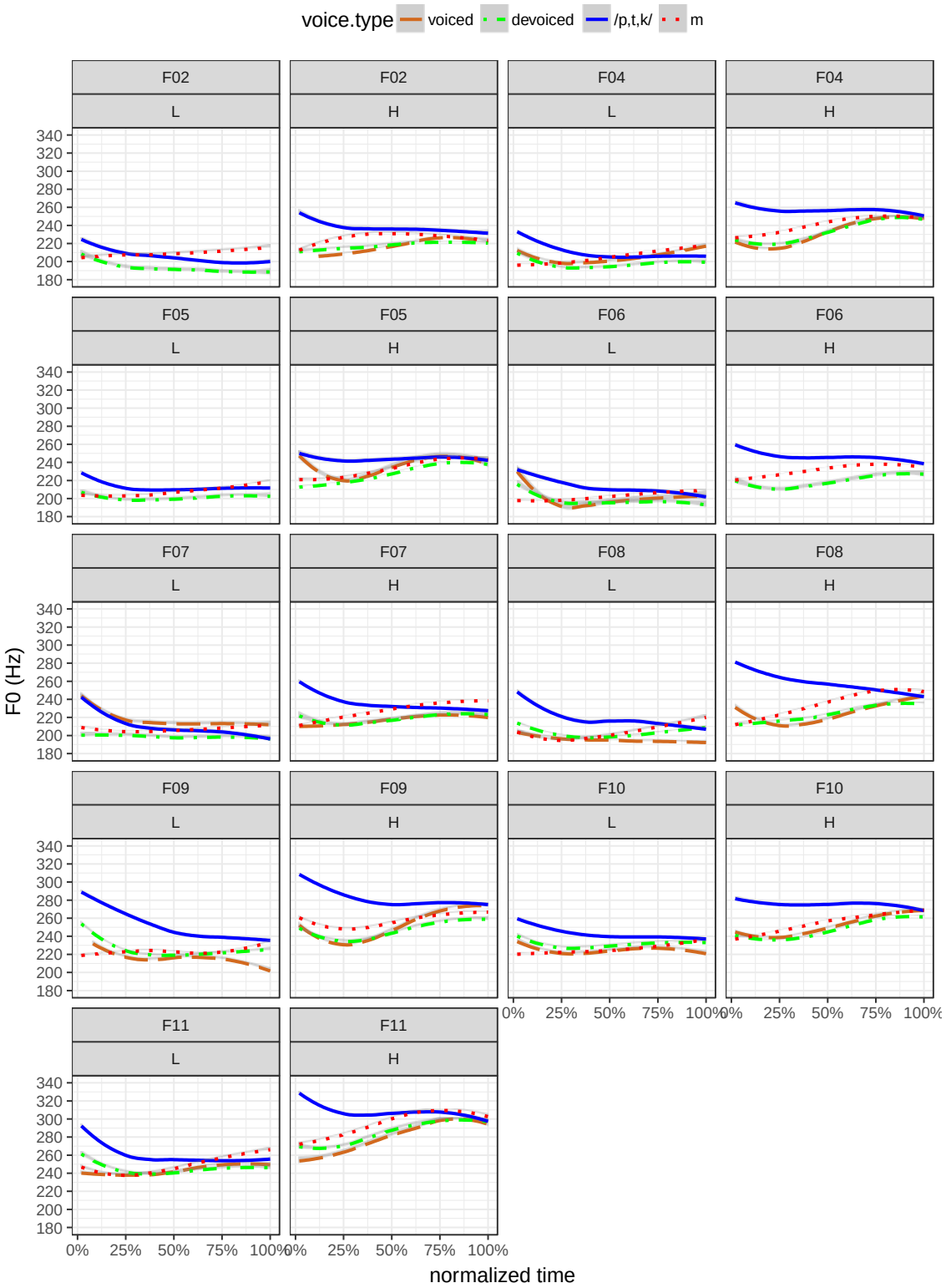
Appendix 3. f0 curves after plosive onsets in word-initial position

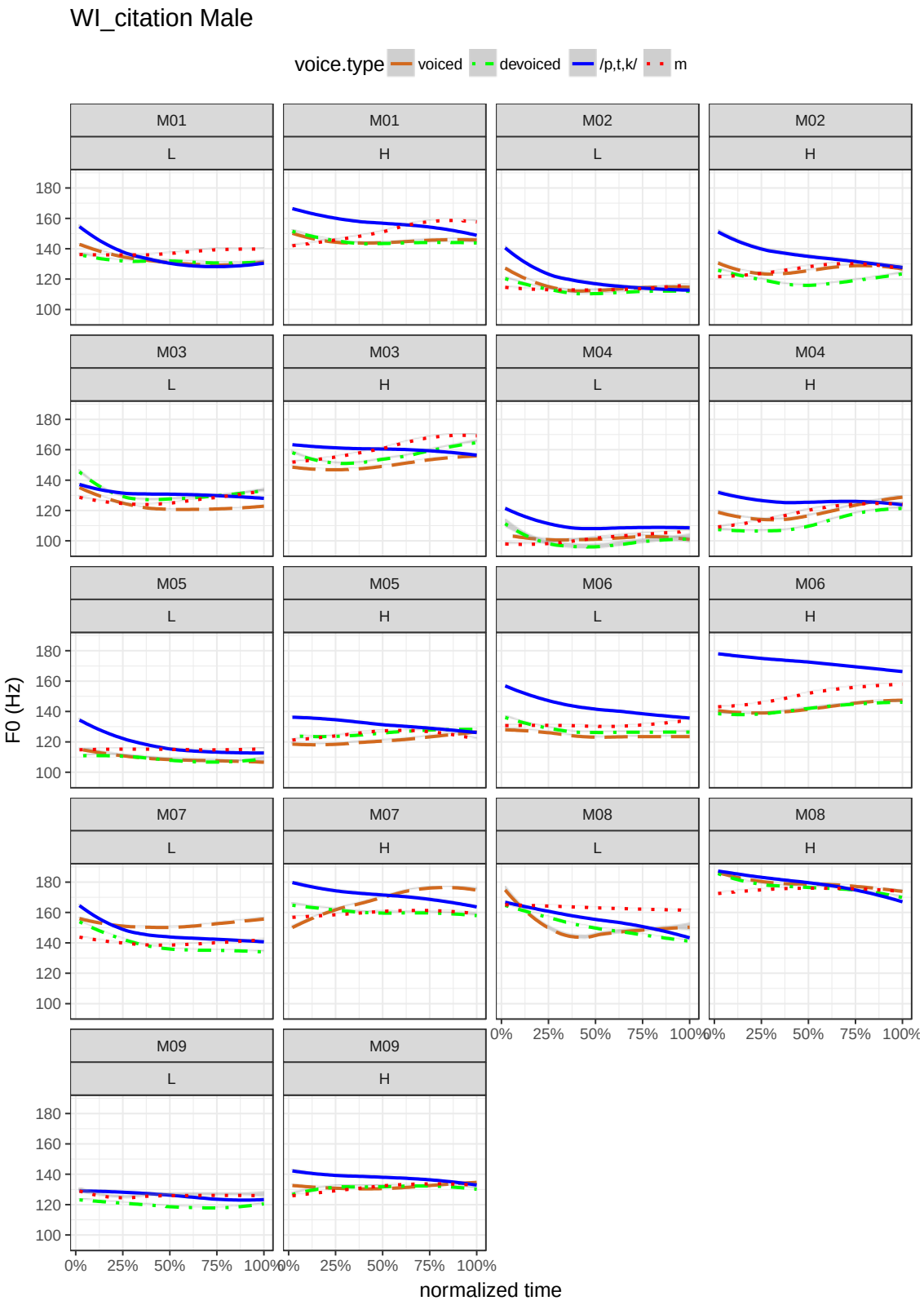



Male
L
H
WI_citation
WI_focus
WI_uttmed
F0 (Hz)
200
175
150
125
100
0%
25%
50%
75%
100%
normalized time
POA
p
t
k
contrast
voiced
voiceless

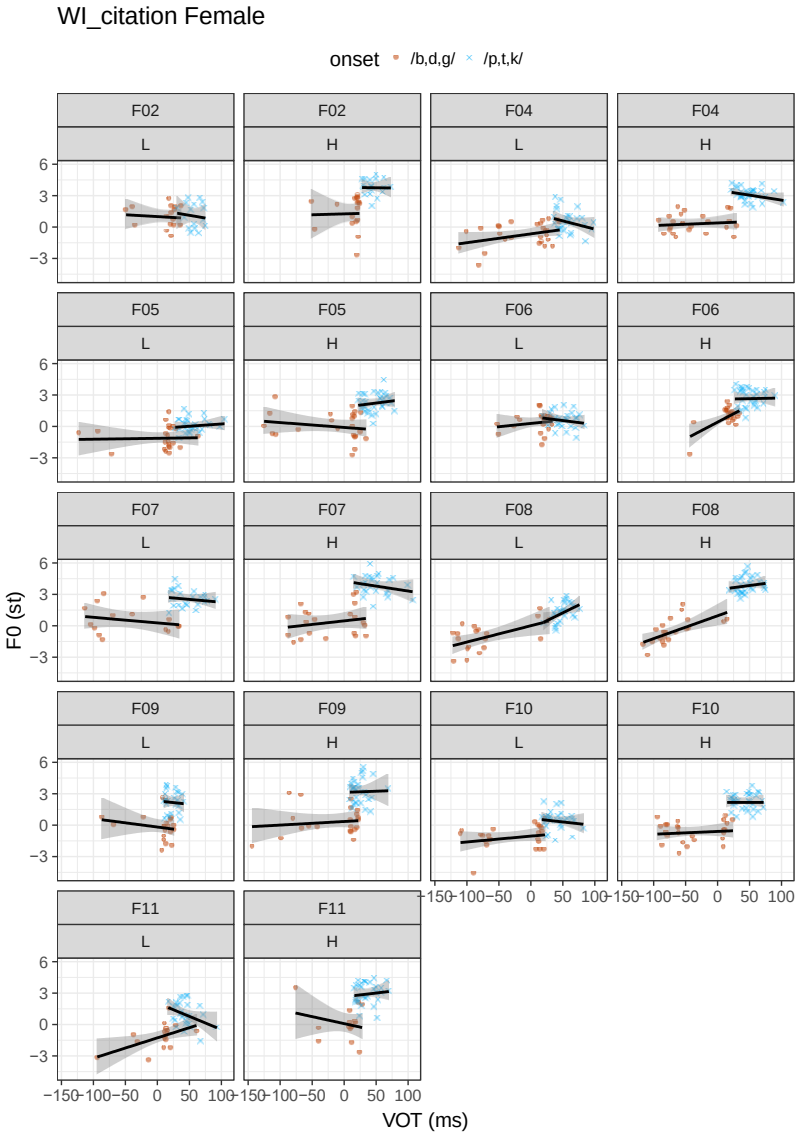
Appendix 4. Individual f0 plots for WI_citation

WI_citation Female

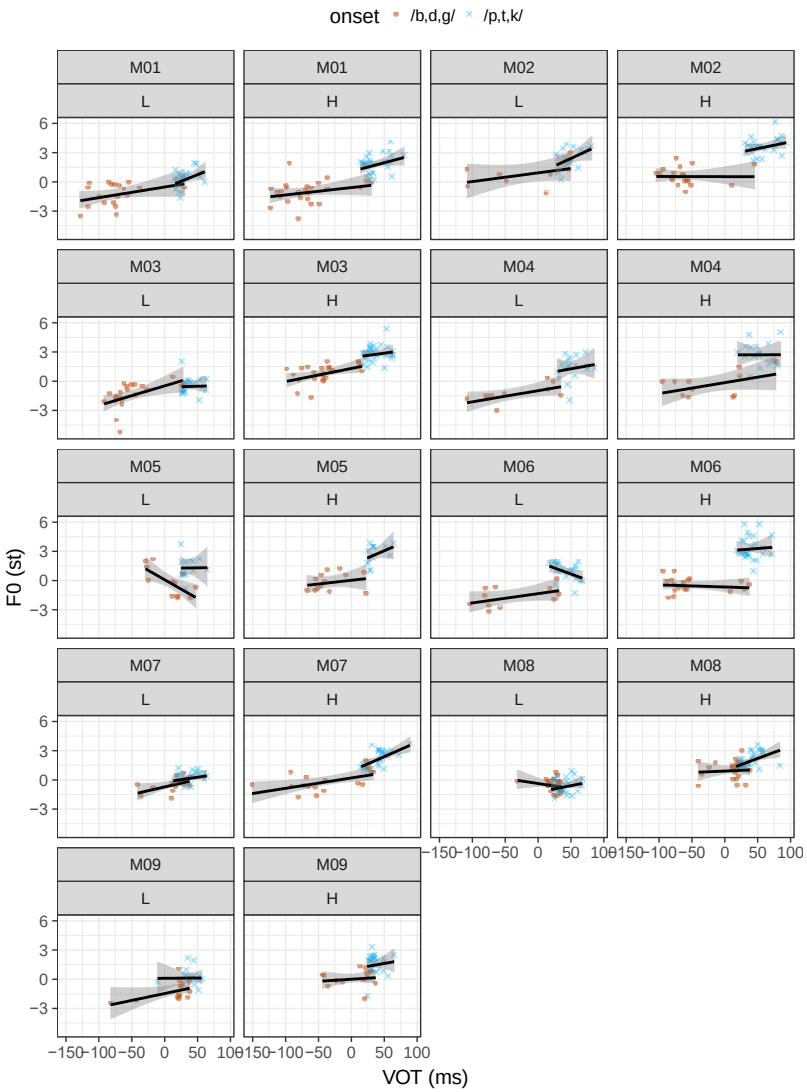



WI_citation Male
voice.type
voiced
devoiced
/p,t,k/
m
M01
L
M01
H
M02
L
M02
H
M03
L
M03
H
M04
L
M04
H
M05
L
M05
H
M06
L
M06
H
M07
L
M07
H
M08
L
M08
H
M09
L
M09
H
F0 (Hz)
180
160
140
120
100
0%
25%
50%
75%
100%
normalized time

Appendix 5. Individual scatter plots of VOT against onset f0 for WI_citation



WI_citation Male



Reference List

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, *63*, 75-86.

van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics*, *32*(4), 455-491.

Amano, S., & Kondo, T. (2000). *NTT database series: Nihongo-no goitokusei [Lexical properties of Japanese]*, 2nd release. Tokyo: Sanseido.

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.

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

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259-284.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255-309.

Boersma, P., & Weenink, D. (1992-2017). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.33.

Caramazza, A., & Yeni-Komshian, G. H. (1974). Voice onset time in two French dialects. *Journal of phonetics*, *2*(3), 239-245.

Chen, Y. (2011). How does phonology guide phonetics in segment–*f0* interaction?. *Journal of Phonetics*, *39*(4), 612-625.

Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, *10*(3), 120-141.

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*(2), 155-190.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics*, *27*(2), 207-229.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121-157.

Cho, T., Whalen, D. H., & Docherty, G. (2019). Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *Journal of Phonetics*, *72*, 52-65.

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.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77-95.

Docherty, G. J. (1992). *The timing of voicing in British English obstruents*. Berlin; New York: Foris.

Erickson, D. (1993). Laryngeal muscle activity in connection with Thai tones. *Research Institute of Logopedics and Phoniatrics Annual Bulletin*, *27*, 135-149.

Ewan, W. G. (1976). *Laryngeal behavior in speech*. Ph.D. dissertation. University of California.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of phonetics*, *29*(2), 109-135.

Gandour, J. (1974). Consonant types and tone in Siamese. *Journal of Phonetics*, *2*, 337-350.

Gordon, M. (2016). Consonant-tone interactions: a phonetic study of four indigenous languages of the Americas. In H. Avelino, M. Coler, & W.L. Wetzels (eds.), *The Phonetics and Phonology of Laryngeal Features in Native American Languages*, 129-156. Leiden; Boston: Brill.

Hagège, C., & Haudricourt, A.-G. (1978). *La Phonologie Panchronique*. Paris: Presses Universitaires de France.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*, 613-617.
Halle, M., & Stevens, K. N. (1971). A note on laryngeal features. *MIT Quarterly Progress Report*, *101*, 198-213.
Hallé, P. A. (1994). Evidence for tone-specific activity of the sternohyoid muscle in Modern Standard Chinese. *Language and Speech*, *37(2)*, 103-123.
Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America*, *125*(1), 425-441.
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.
Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*, 37-58.
Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech*, *42*(4), 401-411.
Hoole, P., & Honda, K. (2011). Automaticity vs. feature-enhancement in the control of segmental f0. In G. N. Clements & R. Ridouane (eds.), *Where do phonological features come from?* 131-171. Amsterdan: John Benjamins.
House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105-113.
Hsu, C.-S., & Jun, S.-A. (1999). Prosodic strengthening in Taiwanese: Syntagmatic or paradigmatic? *UCLA Working Papers in Phonetics*, *96*, 69-89.
Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitch-accent. *Language Sciences*, *31*(2-3), 213-238.
Hyman, L. M. (2011). Tone: Is it different?. In J. Goldsmith, J. Riggle, & A. Yu (eds.), *The Handbook of Phonological Theory*, 2nd Ed.
Hyman, L. M. (2013). Enlarging the scope of phonologization. In A. Yu, (ed.), *Origins of sound change: Approaches to phonologization*, 3-28. Oxford: Oxford University Press.
Itô, J., & Mester, R. A. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic inquiry*, *17*, 49-73.
Iverson, G., & Salmons, J. (1995). Aspiration and laryngeal representations in Germanic. *Phonology*, *12*(3), 369-396.
Jessen, M. (1998). *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins.
Kamiyama, T. (2003). Initial pitch in words beginning with a CVV syllable with a long vowel in Tokyo Japanese. *Proceedings of the 15th International Congress of Phonetic Sciences*, 543-546.
Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of [+ voice] in Japanese. *Language*, *82*, 536-574.
Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, *60*(2), 286-319.
Kim, S., Kim, J., & Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, *71*, 65-80.
Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.
Kirby, J. P. (2018). Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages. *Journal of Phonetics*, *71*, 326-354.
Kirby, J. P., & Ladd, D. R. (2015). Stop voicing and F0 perturbations: Evidence from French and Italian. *Proceedings of the 18th International Congress of Phonetic Sciences*, paper No. 740, 1-5.
Kirby, J. P., & Ladd, D. R. (2016). Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *The Journal of the Acoustical Society of America*, *140*(4), 2400-2411.
Kohler, K. J. (1982). F0 in the production of lenis and fortis plosives. *Phonetica*, *39*(4-5), 199-218.

Kohler, K. J. (1985). F0 in the perception of lenis and fortis plosives. *The Journal of the Acoustical Society of America*, *78*, 21-32.
Kulikov, V. (2012). Voicing and voice assimilation in Russian stops. Ph.D. dissertation. University of Iowa.
Ladd, D. R., & Schmid, S. (2018). Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops. *Journal of Phonetics*, *71*, 229-248.
Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0.
Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *The Journal of the Acoustical Society of America*, *33*(4), 419-425.
Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, *33*(1), 42-49.
Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech*, *29*(1), 3-11.
Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.
Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and speech*, *10*(1), 1-28.
Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America*, *85*(3), 1314-1321.
Lyman, B. S. (1894). The change from surd to sonant in Japanese compounds. *Oriental Club of Philadelphia*, 1-17.
Mikuteit, S., & Reetz, H. (2007). Caught in the ACT: The timing of aspiration and voicing in East Bengali. *Language and Speech*, *50*(2), 247-277.
Nasukawa, K. (2005). The representation of laryngeal-source contrasts in Japanese. In J. van de Weijer, K. Nanjo & T. Nishihara, (eds.), *Voicing in* Japanese, 71-87. Berlin; New York: Mouton de Gruyter.
Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, *24*(1), 1-18.
Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (ed.), *The production of speech*, 189-216. New York: Springer-Verlag.
Ohala, J. J. (2011). Accommodation to the Aerodynamic Voicing Constraint and its phonological relevance. *Proceedings of the 17th International Congress of Phonetic Sciences*, 64-67.
Ohala, J. J., & Riordan, C. (1979). Passive vocal tract enlargement during voiced stops. In J. J. Wolf & D. H. Klatt (eds.), *Speech communication papers*, 89-92. New York: Acoustical Society of America.
Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America*, *75*(1), 224-230.
Pape, D., & Jesus, L. M. (2011). Devoicing of phonologically voiced obstruents: Is European Portuguese different from other Romance languages. *Proceedings of the 17th International Congress of Phonetic Sciences*, 1566-1569.
Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693-703.
Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Docherty and D. R. Ladd (eds.), *Papers in laboratory phonology II: gesture, segment, prosody*, 90-117. Cambridge University Press.
R Core Team (2017). R: A Language and Environment for Statistical Computing. R version 3.4.2. <https://www.R-project.org/>. R Foundation for Statistical Computing.
Raphael, L. J., Tobin, Y., Faber, A., Most, T., Kollia, H. B., & Milstein, D. (1995). Intermediate values of voice onset time. In F. Bell-Berti, L. J. Raphael (eds.), *Producing Speech: Contemporary Issues: for Katherine Safford Harris*, 117-127. New York: AIP Press.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological bulletin*, *92*(1), 81.
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.
Ringen, C., & Kulikov, V. (2012). Voicing in Russian stops: Cross-linguistic implications. *Journal of Slavic Linguistics*, *20(2)*, 269-286.
Serniclaes, W. (1986). *Etude expérimentale de la perception du trait de voisement des occlusives du français.* Ph.D. dissertation. Faculté des Sciences Psychologiques et Pédagogiques.
Shibata, T., & Shibata, R. (1990). Akusento wa doo'ongo o donoteido benbetsu shiuruno ka?: Nihongo, eigo, chuugokugo no baai [Can lexical accent distinguish homonyms in Japanese, English, and Chinese? *Keiryoo Kokugo-gaku [Mathematical Linguistics]*, *17*, 317–237.
Shimizu, K. (1996). *A cross-language study of voicing contrasts of stop consonants in six Asian languages*. Tokyo: Seibido.
Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95-EL101.
Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(2), 287-308.
Silverman, D. (1997). *Phasing and recoverability (Outstanding Dissertations in Linguistics)*. New York: Garland.
Snoeren, N. D., Hallé, P. A., & Segui, J. (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, *34*(2), 241-268.
Solé, M. J. (2018). Articulatory adjustments in initial voiced stops in Spanish, French and English. *Journal of Phonetics*, *66*, 217-241.
Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *The Journal of the Acoustical Society of America*, *62*(2), 435-448.
Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *The Journal of the Acoustical Society of America*, *118*(2), 1026-1037.
Takada, M. (2011). *Nihongo no gotou heisa'on no kenkyuu: VOT no kyoujiteki bunpu to tsuujiteki henka [Research on the word-initial stops of Japanese: Synchronic distribution and diachronic change in VOT]*. Tokyo: Kurosio.
Takada, M., Kong, E. J., Yoneyama, K., & Beckman, M. E. (2015). Loss of prevoicing in modern Japanese /g, d, b/. *Proceedings of the 18th International Congress of Phonetic Sciences*, paper No. 873, 1-5.
Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, NY: SUNY Press.
Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152-2159.
Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of phonetics*, *23*(3), 349-366.
Xu, C. X., & Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association*, *33*(2), 165–181.