More on the articulation of devoiced [u] in Tokyo Japanese: effects of surrounding consonants

Jason A. Shaw[a] & Shigeto Kawahara[b]
[a]Department of Linguistics, Yale University,
New Haven, CT 06520, USA
[b]The Institute of Cultural and Linguistic Studies, Keio University,
Minato-ku, Tokyo 108-8345, Japan

Running head: Articulation of devoiced [u] in Japanese

Corresponding author: jason.shaw@yale.edu, +1 203-432-8289 (J.A. Shaw).

Keywords: Japanese, devoicing, EMA, articulation, deletion, phonetic interpolation, gestural coordination, syllable contact

Abstract

Several aspects of high vowel devoicing in Tokyo Japanese have been extensively studied. One aspect of the phenomenon that remains understudied is the lingual articulation of devoiced vowels, including whether devoiced vowels retain their lingual gesture. Shaw & Kawahara (2018b) addressed this question using EMA (Electro-Magnetic Articulography), finding optional but categorical deletion patterns: some vowels retained a full lingual target, just like their voiced counterparts, whereas other vowels showed trajectories that are best modelled as targetless, i.e., linear interpolation between the surrounding vowels. Extending this finding, as well as being inspired by various phonetic and phonological considerations, the current study explores the hypothesis that this probabilistic deletion of devoiced high vowels may be modulated by the identity of the surrounding consonants. A new follow-up EMA-based experiment with an extended stimulus set replicates the core finding of Shaw & Kawahara (2018b) that Japanese devoiced [u] sometimes lacks a tongue body raising gesture. The current results moreover show that surrounding consonants do indeed affect the probability of tongue dorsum targetlessness. We found that deletion of devoiced vowels is affected by the place of articulation of the preceding consonant, with deletion more likely following a coronal fricative than a labial fricative. Additionally, we found that the manner combination of the flanking consonants, fricative-fricative vs. fricative-stop also has an effect, at least for some speakers; however, unlike the effect of C1 place, the direction of the manner combination effect varies across speakers with some deleting more often in fricative-stop environments and others more often in fricative-fricative environments.

1 Introduction

1.1 General background

Vowels that are adjacent to voiceless obstruents are sometimes produced without vocal fold vibration—i.e. as voiceless—the phenomenon generally referred to as "vowel devoicing." This pattern is observed systematically across many genetically-unrelated languages, including but not limited to Cheyenne (Rhodes, 1972; Vogel, 2021), French (Smith, 2003), Greek (Dauer, 1980), Korean (Jun et al., 1998), (Andean) Spanish (Delforge, 2008), Uspanteko (Bennett, 2020), Uzbek (Sjoberg, 1963), and Turkish (Jannedy, 1995). Tokyo Japanese also exhibits such devoicing of high vowels, and Japanese is arguably the best studied language in this respect (Fujimoto 2015 for a recent review).

A general characterization of the high vowel devoicing pattern in Tokyo Japanese is that high vowels are devoiced between two voiceless obstruents, as well as after a voiceless obstruent and before a pause, although as we will review below, the likelihood of devoicing is affected by various other factors, both linguistic and social. Previous studies have explored this devoicing process from various perspectives, each bearing upon some important issues in phonetic and phonological theory, including how devoicing is implemented in terms of the laryngeal gesture (Fujimoto et al., 2002; Sawashima, 1971; Yoshioka, 1981), how the precise environment affects the likelihood of devoicing (Maekawa & Kikuchi, 2005; Tsuchida, 1997), its categorical or gradient nature (Nielsen, 2015; Tanner et al., 2019), its interaction with lexical pitch accent (Kuriyagawa & Sawashima, 1989; Maekawa, 1990; Vance, 1987) and other prosodic properties (Kilbourn-Ceron & Sonderegger, 2018), its consequences (or lack thereof) for prosodic reorganization (Kondo, 1997, 2001; Kawahara & Shaw, 2018), its perceptual consequences (Cutler et al., 2009; Ogasawara, 2013; Sugito & Hirose, 1988; Whang, 2019), its role in child-directed speech (Fais et al., 2010; Martin et al., 2014), its acquisition patterns (Imaizumi & Hayashi, 1995; Imaizumi et al., 1995) as well as the influence of social factors on this pattern (Imai, 2004; Imaizumi et al., 1995). There is no doubt that these studies have revealed important aspects of this devoicing phenomenon in Japanese, and we understand its nature much better than 50 years ago.

However, despite the accumulation of studies on vowel devoicing, one aspect that is heavily under-addressed regarding the pattern of high vowel devoicing in Japanese—and any other languages that exhibit vowel devoicing, for that matter—is the question of how the lingual gesture is implemented for the devoiced vowels. This issue is related to the question of whether these devoiced vowels are phonologically deleted or not; if the high vowels are phonologically deleted, then we would expect them to lack any lingual gesture. If the process at issue is phonologically devoicing rather than deletion, on the other hand, we may expect that their lingual gestures are retained. Vance (2008), which is the most recent and updated textbook on Japanese phonetics and phonology, indicates that this issue is not settled. The current paper reports an experiment that addresses this issue by analyzing articulatory kinematics via Electromagnetic Articulography. This paper moreover tests a new, specific hypothesis that deletion probability may be modulated by the surrounding consonantal environment. We will start with the overview of the relevant literature on this topic, which leads us to examine this specific hypothesis.

1.2 Are devoiced vowels in Japanese deleted?

Since devoiced vowels lack a periodic energy source, it is difficult, if not entirely impossible, to infer from their acoustic profiles alone whether devoiced vowels retain their lingual gestures or not. There are some studies which addressed this question via impressionistic observations. Kawakami (1977) argues that vowels delete completely in some environments but not others, but he offers no experimental evidence for this claim. Vance (1987) examines the hypothesis that devoiced high vowels in Japanese are entirely deleted but ultimately rejects this hypothesis. Kondo (1997, 2001) argues that devoiced high vowels are deleted based on a phonological consideration: vowel devoicing in consecutive syllables is often inhibited (though see Nielsen 2015), and Kondo (1997, 2001) attributes this observation to a constraint against triconsonantal clusters. The underlying assumption of this analysis is that devoiced vowels are deleted, resulting in consonant clusters.

On the other hand, Tsuchida (1997) and Kawahara (2015) point out that devoiced vowels count toward a bimoraic requirement in foot-based morphophonological truncation patterns (Poser, 1990), arguing that these vowels do not delete phonologically. Like these two authors, Hirayama (2009) demonstrates that devoiced vowels behave just like voiced vowels in the Japanese *haiku* poetry pattern, which is mora-based (Vance, 1987).

In line with this view, Jun and her colleagues advanced an explanation of high vowel devoicing (in Korean) in terms of "gestural overlap" (Browman & Goldstein, 1992a), according to which the articulatory gesture of high vowels is overlapped in time by the laryngeal glottal abduction gesture of surrounding consonants (Jun & Beckman, 1993; Jun et al., 1998). In this gestural overlap view, Japanese phonology does not delete the devoiced high vowels; the high vowels are merely rendered inaudible because of the glottal abduction gesture that coincides in time with the vocalic gesture. This is analogous to the famous case of English *perfect memory*, in which the word-final [t] in *perfect* can be made inaudible due to gestural overlap with the following [m], even when the [t]'s tongue tip gesture remains intact (Browman & Goldstein, 1992a). Similar to this case in English, it is conceivable that lingual gestures of devoiced high vowels are present, but are merely rendered inaudible because of the overlapping glottal abduction gesture. In this gestural-overlap scenario, it is also possible that lingual gestures are reduced, rather than remaining completely intact, assuming speakers invest less articulatory energy into sounds that are difficult to perceive and hence may not contribute much to lexical retrieval (e.g. Hall et al. 2018; Jaeger & Buz 2018).

Recently we have witnessed a rise of studies addressing this question—whether devoiced high vowels are deleted or not—using instrumental techniques. Beckman and her colleagues, based on the inspection of spectrograms, argue that devoiced vowels are physically not present (Beckman, 1982; Beckman & Shoji, 1984), suggesting that the pattern should be characterized as deletion, although they also suggest that it may make sense to characterize the pattern as devoicing, not deletion, from the psycholinguistic perspective; i.e. Japanese speakers feel that "vowels are there" even when they are actually deleted (cf. Dupoux et al. 1999, 2011; Whang 2019). This is possibly because of coarticulatory influences of vowels on flanking consonants that remain even when typical acoustic cues to the vowel are absent. It is known, for example, that consonant identity influences vowel quality in perceptual epenthesis (Durvasula et al., 2018; Kilpatrick et al., 2020).

Faber & Vance (2010) offer some acoustic evidence for the hypothesis that vowel devoicing is best characterized as gestural overlap of laryngeal gestures in Japanese (Jun & Beckman, 1993; Jun et al., 1998). Jannedy (1995) and Bennett (2020) entertain a similar hypothesis for devoiced vowels in Turkish and Uspanteko, respectively. Whang (2018) measured COG during devoiced vowels in Japanese and argues that some devoiced vowels in Japanese are in fact deleted, while others are not. More specifically, Whang (2018) argues that deletion is more likely in the environment where the quality of those vowels can be recovered from surrounding consonants; e.g. after [F], only [u][1] is possible, while

[1] Here and throughout the paper, we use the IPA symbol [u] to denote the high non-front vowel in Japanese. The exact phonetic nature of this vowel, as well as how to transcribe it, is a contentious issue

after [ɕ] both [u] and [i] are possible (see also Whang 2019).

However, generally speaking, there are limits on how much we can conclude about the articulatory gestures from their resulting acoustic signals (see e.g. Browman & Goldstein 1989; Guenther et al. 1999; Munson et al. 2010; Perkell et al. 1993). It is thus important that we address the nature of the lingual gesture of devoiced high vowels through observation of articulatory movement. To that end, Matsui (2017) used EPG (Electro PalatoGraphy) to examine the linguo-palatal contact pattern of devoiced syllable [sɯ̥], and showed that the pattern remains very constant across the syllable; i.e. there does not seem to be a clear change in the linguo-palatal contact pattern from [s] to [ɯ̥], implying that the lingual gesture of the devoiced [ɯ̥] is absent. Nakamura (2003) on the other hand reports that vestiges of lingual gestures of devoiced vowels can be found in his EPG data. Although these two results, which seem to conflict with each other, are telling, there are limits on how much we can conclude about tongue body movement—primary correlates of vowel gestures (Browman & Goldstein, 1992b; Johnson et al., 1993)—from EPG data in general, since EPG only registers contact with the palate. Funatsu & Fujimoto (2011) used ElectroMagnetic Midsagittal Articulography (EMMA) to study articulatory gestures of devoiced [i], showing that the articulatory gesture of [i] is comparable between voiced [i] and devoiced [i̥]. This study however used one speaker and one pair of items (/kide/ vs. /kite/) with four repetitions, and offers no quantitative comparisons between the two voicing conditions.

The most extensive study on this topic—the presence/absence of lingual gestures of devoiced vowels in Japanese—to date is that of Shaw & Kawahara (2018b), who used EMA (ElectroMagnetic Articulography) to study the articulatory nature of devoiced [ɯ̥]s of six naive speakers of Tokyo Japanese, and the current paper can be considered as a direct follow-up of Shaw & Kawahara (2018b).

Shaw & Kawahara (2018b) analyzed four dyads to compare the articulatory trajectories of CVC sequences, in which one member of each dyad contains a voiced vowel and the other a devoiced vowel. The four dyads were: (1) [ɸɯ̥soku] vs. [ɸɯzoku], (2) [ɕɯ̥taisee] vs.

even in the contemporary literature (Vance, 2008). We will return to this issue in the method section (§3.2), where we justify our choice of phonetic parameters used to assess the deletion of this vowel.

[ɕudaika], (3) [katsu̥toki] vs. [katsudoo] and (4) [masu̥taa] vs. [masuda].[2] Their strategy, reviewed in further detail below in §3.2, is to compare the articulatory trajectory of [Cu̥C] with respect to that of [CuC] and [C;C], the latter of which is characterized by linear interpolation between the surrounding vowels (Choi, 1995; Cohn, 1993; Keating, 1988; Pierrehumbert & Beckman, 1988). Their conclusion in a nutshell is that some productions contain no articulatory target, while others show lingual targets that are no different from voiced vowels; i.e. they found a pattern of optional but categorical deletion. Moreover, they found some variation with respect to how often each item showed devoiced vowels without lingual targets: devoiced vowels were more likely to be targetless between [ɕ] and [t] ([ɕu̥taisee]) than between [F] and [s] ([Fu̥soku]). This asymmetry was consistent across the speakers (see also discussion in §3.2).

An intriguing hypothesis that emerges from this result is that vowel deletion probability may be systematically modulated via surrounding consonant environments—Japanese [u̥]s may be more likely to lack a lingual gesture between a fricative and a stop than between two fricatives. We expand in the next subsection why this is an interesting and plausible hypothesis to entertain, although we also note at this point that the results by Shaw & Kawahara (2018b) are based on just one dyad per each phonological condition.

1.3 The current hypothesis

The general hypothesis pursued in this study is that the probability of [u̥] lacking its lingual gesture—which we equate with the probability of phonological deletion for the sake of exposition here (see Shaw & Kawahara 2018b)—is modulated by surrounding consonantal environment. A more specific hypothesis is that [u̥]s are more likely to be phonologically deleted when surrounded by a fricative and a stop than when surrounded by two fricatives. As mentioned above, one reason to entertain this hypothesis is the results reported by Shaw & Kawahara (2018b), who show that targetless [u̥]s were more likely in [ɕu̥taisee] than in [Fu̥soku]. However, it is hard to know whether or not their findings are generalizable to other items with similar phonological properties, because their results are based on one

[2] Glosses: (1) shortage vs. attachment, (2) subjectivity vs. theme song, (3) when winning vs. activity and (4) master vs. PERSONAL NAME.

dyad per each phonological condition.

Nevertheless, this hypothesis dovetails with an observation by Starr & Shih (2017), who found that devoiced vowels are often skipped in the text-setting of Japanese songs, and this is especially so when they are surrounded by a fricative and a stop. Their observation may suggest that Japanese composers are sensitive to the higher likelihood of vowel deletion in this environment. The higher likelihood of deletion after a fricative and before a stop is also compatible with the general cross-linguistic observation about prosodic wellformedness, namely, syllable contact laws (Murray & Vennemann, 1983; Vennemann, 1988)—languages generally prefer sonority fall to sonority plateau/rise across a syllable boundary. To the extent that Japanese is also sensitive to such prosodic wellformedness conditions (cf. Berent et al. 2007, 2008), we may expect Japanese high vowels to delete more often in the environment which would result in a fricative-stop cluster than a fricative-fricative cluster. To view it from the opposite perspective, if it can be shown that Japanese speakers delete high vowels in accordance with syllable contact law, it may imply that speakers of Japanese, generally considered to be a "CV-language" disallowing hetero-organic consonant clusters, are sensitive to wellformedness conditions on consonant clusters (see Berent et al. 2007, 2008 for related ideas), possibly because they can extrapolate sonority-based patterns from limited data (Daland et al., 2011).

There are other reasons to entertain the current hypothesis. Previous studies have shown that devoicing (not necessarily deletion) is more likely between a fricative and a stop than between two fricatives (see e.g. Fujimoto 2015; Hirayama 2009; Maekawa & Kikuchi 2005; Martin et al. 2014; Tsuchida 1997). Let us suppose that devoicing and deletion are on the same "reduction continuum".[3] Then, everything else being equal, we may expect deletion to be more likely in the environment where devoicing is more likely in the first place. This leads us to expect that deletion is more likely between a fricative and a stop, because devoicing is more likely in this environment.

A recent acoustic study by Whang et al. (2020) suggests that devoicing and deletion

[3] In Kagoshima Japanese, word-final high vowels—those that are devoiced—undergo phonological deletion, which feeds other phonological changes of consonants in the word-final syllables (Haraguchi, 1984; Kaneko & Kawahara, 2002; Kibe, 2001). It seems natural to consider deletion as the extreme end of the reduction continuum, and that devoicing is one step in the continuum before deletion (see Haraguchi 1984; McCarthy 2008; Tsuchida 1997).

may both be characterized as enhancement strategies of the larynx abduction gesture. Fujimoto et al. (2002) as well as Sawashima (1971) show that devoiced vowels in Japanese involve an active abduction gesture, and thus there is a sense in which speakers are actively signalling "voicelessness." According to Whang et al. (2020), vowel devoicing in fact raises COG of the aperidoc energy of surrounding obstruents, possibly due to wider glottal aperture and increased airflow, and deletion of the tongue dorsum raising gesture for [u] further raises that COG. This hypothesis too leads us to expect that devoicing and deletion should work in tandem with each other, as deletion can enhance the auditory cue to devoicing. To the degree that devoicing is more likely after a fricative and before a stop than between two fricatives (see above), deletion may show the same probabilistic pattern.

All of these considerations—prosodic wellformedness, reduction, enhancement of devoicing—converge on the same prediction: deletion should be more likely when it results in a fricative-stop sequence than when it results in a fricative-fricative sequence. Existing evidence from Shaw & Kawahara (2018b) is consistent with this conclusion; however, the evidence to date is rather thin.

To recap, the current experiment was set out to examine the general hypothesis that vowel deletion probability is modulated by surrounding consonants. The more specific hypothesis is that deletion is most likely between a fricative and a stop, and less likely between two fricatives. The experiment also serves as an attempt to replicated the basic findings of Shaw & Kawahara (2018b)—devoiced [u̥]s in Japanese are optionally deleted—with a much extended set of stimuli.

2 Experimental methods

The current experiment measured and analyzed the tongue dorsum trajectories of devoiced [u̥], using EMA (Electro Magnetic Articulatograph). Most of the experimental details follow those of Shaw & Kawahara (2018b). One distinct characteristic of this approach that we would like to highlight at this stage is that it assesses the presence of an articulatory target on a token-by-token basis, rather than analyzing averaged contours. This strategy is important because analyzing averaged contours cannot distinguish two different phonological hypotheses: reduction vs. optional deletion (Cohn, 2006; Kawahara et al., to appear;

Shaw & Kawahara, 2018a). Lingual gestures of devoiced vowels, even when not phonologically deleted, are conceivably reduced in magnitude, since the vowel gestures are not as audible due to devoicing and do not contribute much to lexical access anyway (e.g. Hall et al. 2018; Jaeger & Buz 2018; see also Lindblom 1990). Interpreting any difference between voiced vowels and devoiced vowels as deletion would therefore be hasty.

On the other hand, as Shaw & Kawahara (2018b) found, devoiced vowels can retain their full lingual gestures, showing comparable movement trajectories to voiced vowels, but they can also sometimes be deleted. Averaging over cases of full targets and cases of categorical deletion can lead to an erroneous conclusion that the overall pattern supports the reduction hypothesis (Cohn, 2006).

This specific problem can be illustrated by a comparison of two recent studies. Kawahara et al. (to appear) developed a token-by-token analysis of the F0 patterns of the dataset recorded and analyzed by Ishihara (2011). The averaged-based analysis by the latter concluded that pitch accent after wh-elements in Japanese is reduced. On the other hand, a token-by-token reanalysis by Kawahara et al. (to appear) shows that at least some speakers show a mixture of full target and deletion. This comparison shows that when both deletion and reduction are theoretically-justifiable hypotheses, it is important that we distinguish between them through a token-by-token analysis.

In addition to avoiding this general problem of resorting to average-based analyses, the current analysis has a virtue of analyzing the entire articulatory trajectories; in the current analysis, no aspects of speech signals wihtin the analysis window are given special status, eschewing the potential danger of missing important aspects of dynamic speech (Cho, 2016; Mücke et al., 2014; Vatikiotis-Bateson et al., 2014).

2.1 Participants

Seven native speakers of Tokyo Japanese (4 male) participated in the current experiment. They were all born in Tokyo, lived there at the time of their participation in the study, and had spent no more than 3 months outside of the Tokyo region. Procedures were explained to participants in Japanese by a research assistant, who was also a native speaker of Tokyo Japanese. All participants were naive to the purpose of the experiment. Participants were compensated for their time and local travel expenses. Data from one speaker had to be excluded, because we were unable to record as many repetitions as other speakers. This speaker was originally coded as Speaker 6; their data is not discussed further below. No speakers who participated in Shaw & Kawahara (2018b) participated in this study, since one of the aims was to examine whether the results of Shaw & Kawahara (2018b) can be generalized to other speakers.

2.2 Stimuli

Following Shaw & Kawahara (2018b), the major target of our analysis is tongue dorsum height in the trajectory of $V_1C_1V_2C_2V_3$ sequence, in which V_2 represents the devoiced vowels in question—justification of this analytical choice is offered below in §3.2. The primary question is whether we would observe a clear rise in tongue dorsum height from V_1 to V_2 and a fall from V_2 to V_3. V_3 was therefore always a non-high vowel in our stimuli. The target vowels (V_2=[u]s) were always word-initial, and V_1 was the last vowel of the preceding word in the frame sentence, [e].

At the time of stimulus design, four conditions were included in order to thoroughly explore the effects of surrounding consonant types: fricative-stop (FS), fricative-fricative (FF), stop-stop (SS), and stop-fricative (SF), consisting of 18 dyads shown in Table 1. All the stimuli were existing words in Japanese, where the members on the left were expected to undergo devoicing. Each dyad constituted near minimal pairs, in which one member contained C_1VC_2 sequence where both consonants are voiceless and the other member contained a minimally different C_1VC_2 sequence in which C_2 is voiced, hence V is not expected to devoice.

Table 1: The list of stimuli recorded in the EMA experiment. S=Stop; F=Fricative. See footnote 5 for glosses. Accent is shown by a following aphostrophy.

FS	FF
/Futon/ vs. /Fudou/	/Fusoku/ vs. /Fuzoku/
/Futan/ vs. /Fu'dan/	/Fusai/ vs. /Fuzai/
/Futa/ vs. /Fuda/	/Fusagaru/ vs. /Fuzake'ru/
/Sutaisei/ vs. /Suda'ika/	/Susai/ vs. /Suzai/
/Sutou/ vs. /Sudou/	/Su'sa/ vs. /Su'zan/
/Sutokou/ vs. /Sudo'uken/	/Su'so/ vs. /Suzou/
SS	**SF**
/kutakuta/ vs. /kudaranu/	/kusami/ vs. /kuzai/
/kutaba'ru/ vs. /kudasa'ru/	/kusari/ vs. /kuzawa/
/kutanijaki/ vs. /kuda'nSita/	/kusaka'ri/ vs. /kuzakitSo/

Choosing existing words with the appropriate segmental compositions did not enable us to control for accentedness within each pair. For example, /Futan/ is unaccented, whereas /Fu'dan/ is accented on the initial syllable. However, Tsuchida (1997) and Martin et al. (2014) show that accent placement has little effect on devoicing patterns among contemporary speakers of Japanese. Durational differences between accented and unaccented syllables are minimal in Japanese (Beckman, 1986), which if substantial, may affect devoicability/deletablity. For these reasons, we judged this difference to be non-crucial.[4]

The current study focused on [u] instead of examining both [u] and [i], both of which are known to devoice. This is partly because the current study is a direct follow-up of Shaw & Kawahara (2018b), who also examined only [u], and also because we needed enough repetitions to execute the computational analysis that was planned (see §3.2 for details). Examining the lingual gesture of devoiced [i] warrants a new set of studies.

After the recording, we came to the conclusion that the conditions in which C_1 is a stop

[4] Shaw & Kawahara (2018b) did not perfectly control for accent between two members within a dyad either, although in their design, [u] is either accented or unaccented within each dyad, i.e. no direct comparison was made between accented [u] and unaccented [u].

(=the SS and SF conditions) could not be reliably analyzed for the following reason. At the time of stimulus design, we decided that C_1 had to be [k], because [p] is not allowed in the native vocabulary (Ito & Mester, 1995), and [t] is affricated before high vowels (Vance, 2008). However, since we were interested in the tongue dorsum height of (devoiced) [u], it was not possible to objectively discern control of tongue dorsum height associated with [k] from tongue dorsum height associated with [u]. For this reasons, this paper focuses on the comparison between FS condition and the FF condition.[5]

2.3 Procedure

Each participant produced 14-15 repetitions of the 36 target words in the carrier phrase: "okkee X to itte" (Ok, say X), where X is a stimulus word. Participants were instructed to speak as if they were making a request of a friend. This was to ensure that the speakers did not speak too formally or too slowly, which may inhibit vowel devoicing in the first place.

This resulted in a corpus of 3,204 tokens (14 or 15 repetitions ð 36 words ð 6 speakers). Words were presented in Japanese script (composed of hiragana, katakana and kanji characters as required for natural presentation) and fully randomized.

2.4 Equipment

We used an NDI Wave ElectroMagnetic Articulograph system sampling at 100 Hz to capture articulatory movement. NDI wave 5DoF sensors (receiver coils) were attached to three locations on the sagittal midline of the tongue, and on the lips, jaw (below the lower incisor), nasion and left/right mastoids. The most anterior sensor on the tongue, henceforth TT, was attached less than one cm from the tongue tip (see Figure 1). The most posterior sensor, henceforth TD, was attached as far back as was comfortable for the participant. A third sensor, henceforth TB, was placed on the tongue body roughly equidistant between the TT and TD sensors. Sensors were attached with attached with a combination of sur-

[5] The glosses for the items that were analyzed are as follows. FF: blanket vs. not moving, burden vs. usual, top vs. amulet, subjectivity vs. main theme, FOOD NAME vs. hand-moving, Tokyo Highway vs. lead; FS: shortage vs. attachment, debt vs. absence, filled vs. joke, organize vs. research, chair vs. abacus, main claim vs. *sake*-making.

gical glue and ketac dental adhesive. Acoustic data were recorded simultaneously at 22 KHz with a Schoeps MK 41S supercardioid microphone (with Schoeps CMC 6 Ug power module).

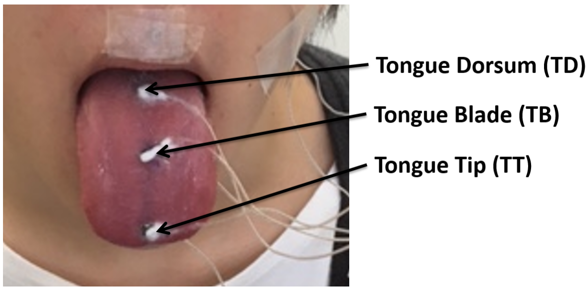

Figure 1: Illustration of the sensor placement (reproduced from Shaw & Kawahara 2018b).

2.5 Stimulus display

Words were displayed on a monitor positioned 25cm outside of the NDI Wave magnetic field. Stimulus display was controlled manually using an Eprime script. This setup allowed for online monitoring of hesitations, mispronunciations and disfluencies. These were rare, but when they occurred, items were marked for repeated presentation by the experimenter. These items were then re-inserted into the random presentation of remaining items. This method ensured that we recorded at least 14 fluent tokens of each target item.

2.6 Post-processing

Following the main recording session, we also recorded the bite plane of each participant by having them hold a rigid object, with three 5DoF sensors attached to it, between their teeth. Head movements were corrected computationally after data collection with reference to three sensors on the head, the left/right mastoid and nasion sensors, and the three sensors on the bite plane. The head corrected data was rotated so that the origin of the spatial coordinates corresponds to the occlusal plane at the front teeth.

3 Data analysis

3.1 Data processing

The wav files recorded in the experiment were submitted to forced alignment, using FAVE.[6] Textgrids from forced alignment were hand-corrected and, during this process, the target vowels were coded for voicing. Most vowels in devoicing environments were in fact devoiced, as evident from visual inspection of the spectrogram and waveform. However, some tokens in the devoicing environment exceptionally retained clear signs of glottal vibration. These vowels were coded as voiced, and excluded from the following computational analysis. The supplementary materials, available at DOI 10.17605/OSF.IO/PGRVZ, provide example spectrograms of voiced and devoiced tokens and a list of all exclusions.

Articulatory data corresponding to each token were extracted based on the textgrids. The data were smoothed using the robust smoothing algorithm (Garcia, 2010) and, subsequently, visualized in MVIEW, a Matlab-based program to analyze articulatory data (Tiede, 2005). Within MVIEW, the consonant gestures flanking the target vowel were parsed using the `findgest` algorithm. `Findgest` identifies gestures semi-automatically based upon the velocity signal in the movement toward and away from gestural targets. An illustrative example is provided in Figure 2. The consonant gestures were used to define a temporal interval for further analysis.[7] Tokens with missing data in the target interval were excluded from further analysis. Some tokens had velocity peaks that were not large enough to clearly parse out movement related to the consonants. If a token was missing a gesture parse for either consonant, it was excluded from further analysis. A total of 239 tokens were excluded for this reason. The resulting data set consisted of 2,431 tokens for analysis, which had clearly distinguishable consonantal gestures flanking the target vowel.

[6] `https://github.com/JoFrhwld/FAVE/wiki/Using-FAVE-align`

[7] The onset of movement of the consonants occurs at a similar time as the maximum tongue height of the preceding vowel. We choose to define the temporal interval for analysis based on the onset of consonant movement instead of,e.g., the maximum TD height in the vicinity of the consonant, primarily because the results presented here are situated in a bigger project which includes also how the reduction/deletion of vowels influences the coordination of flanking consonants.

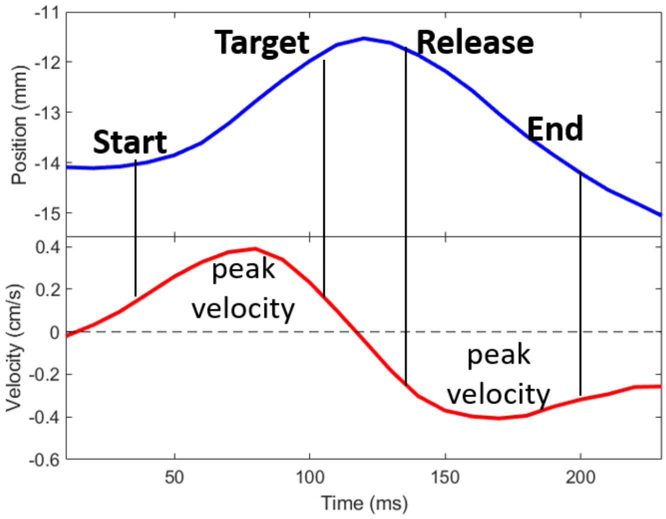

Figure 2: A sample articulatory trajectory and how the articulatory landmarks were identified using findgest.

3.2 Computational analyses

The temporal interval spanning from the onset of movement of C_1, the consonant preceding the target vowel, and the offset of movement of C_2, the consonant following the target vowel, was subjected to further analysis. To address the question of whether devoiced [u̥] has an articulatory target, we focused on tongue height, instead of tongue retraction or lip gestures, both of which have been questioned as reliable articulatory correlates of this vowel in contemporary Japanese (Isomura, 2009; Nogita et al., 2013; Shaw & Kawahara, 2018a; Vance, 2008). Like Shaw & Kawahara (2018b), the analysis focused on the movements of the TD sensor (see Figure 1), the most posterior sensor on the tongue, which is typically used to detect vowel gestures (Browman & Goldstein, 1992b; Johnson et al., 1993).

Figure 3 shows sample trajectories of a voiced vowel (left), a devoiced vowel with a clear tongue dorsum raising during [u] (middle), and a devoiced vowel without a very clear movement in terms of tongue dorsum height (right). The top panels show the audio signal. The second panels from the top show tongue dorsum articulatory trajectories, which are the primary target of our analyses. For reference the third and fourth panels show trajectories related to the flanking consonants. The token in the right panel does not appear to have a clear tongue dorsum raising gesture during [u̥], whereas the [u̥] token in the middle panel does seem to have a clear raising gesture. The challenge is to go beyond such impressionistic classifications and to establish an objective method to classify whether devoiced vowels show a tongue dorsum raising gesture or not.

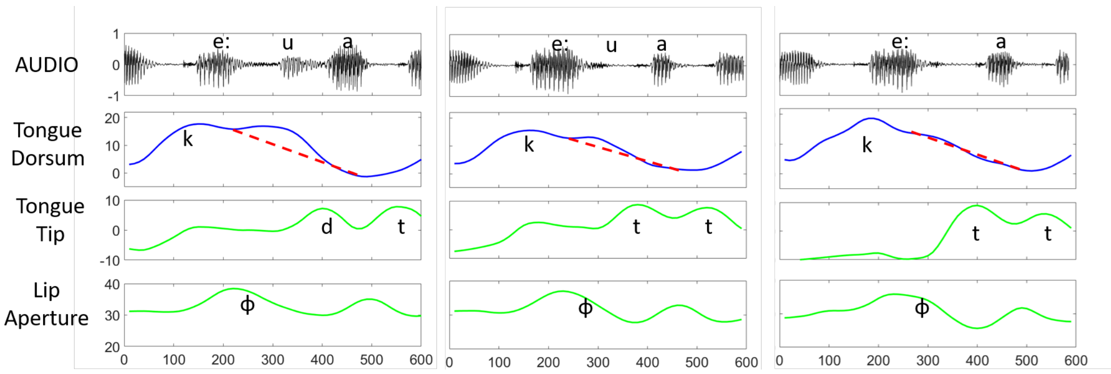

Figure 3: Sample EMA trajectories. The top panels show audio signals. The second panels show the tongue dorsum movement. The dotted red line is a linear interpolation from the preceding vowel to the following vowel.

To do so, we applied the approach described and motivated in detail in Shaw & Kawahara (2018a,b), schematically illustrated in Figure 4. This computational methodology was developed to assess the presence/absence of a lingual vowel target of devoiced vowels in articulatory trajectories. The approach is general enough that it has been extended to other types of continuous phonetic data, including nasal reduction in Ende (Brickhouse & Lindsey, 2020), pitch accent eradication in Japanese (Kawahara et al., to appear), and tone reduction in Mandarin Chinese (Zhang et al., 2019).

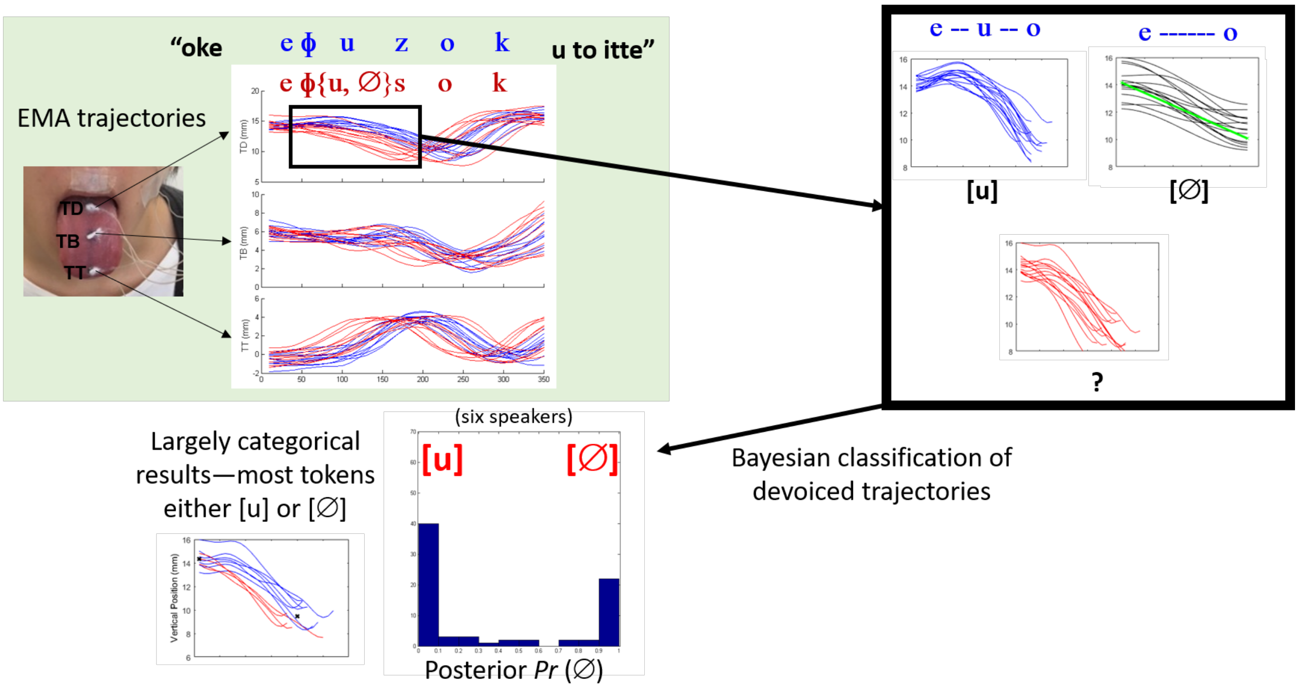

Figure 4: Summary of simulation and classification procedure developped and defended in Shaw & Kawahara (2018b).

The target interval spans from the preceding vowel to the following vowel (see the left upper panel of Figure 4). For example, for the word [Fuzoku], the analysis window starts from [e] in the carrier sentence and includes the main target CVC ([Fuz]) and the following vowel [o]. The question of interest is whether given the vowel sequence [e]-[u]-[o], we would observe a tongue dorsum raising gesture, when [u] is devoiced. When [u]'s tongue dorsum gesture is undoubtedly present, as in the case for voiced [u], we should observe a clear raising gesture (the left panel of Figure 3). On the other hand, if the vowel gesture is deleted, we expect articulatory trajectories that interpolate between [e] and [o] (represented as a green straight line in the right upper panel of Figure 4). Since articulatory movements, as behavioral data more generally, are always noisy actuations of intentions, the challenge is to develop an objective method with which we can assess whether each articulatory contour of a devoiced [u̥] is better characterized as target-present or target-absent (the upper right box in Figure 4). The computational toolkit developed by Shaw & Kawahara (2018a,b) allows us to address this question on a token-by-token basis.

The first step in this computational method is to analyze the articulatory trajectories in a low-dimensional space, by making use of Discrete Cosine Transform (DCT) (e.g. Jain 1989). Through DCT, a signal is transformed into the sum of cosine components of gradually increasing frequency. This transformation is similar to Fourier transform in that timeseries data—here, the articulatory trajectory—is represented in frequency space, i.e., as cosines of varying frequency and magnitude. Unlike Fourier transform, DCT uses only cosines instead of a combination of sines and cosines and there is no imaginary component. Additionally, DCT has compression properties (Jain, 1989), like Principal Component Analysis (PCA)—the articulatory trajectory within the analysis window can often be represented with a small number of DCT components. Because speech articulators are relatively slow, high frequency components are not needed to represent their controlled movement, a point which we demonstrate below.

The numerical expression of DCT is provided in Equations (1) and (2): n is the positional signal, L is the length of the window (in samples), k is the number of the DCT coefficient, which ranges from 1 to L, y is the magnitude of each coefficient, and w is a weight. DCT coefficients can be positive or negative and their absolute value represents the magnitude of their contribution to spatial modulation of the signal. For the first DCT coefficient, the numerator in the scope of the cosine is zero, which means that it equals 1 for every sample n in the trajectory. These are summed, and when multiplied by the relevant weight ($\sqrt{\frac{1}{L}}$), they yield a quantity that is related to the average of the trajectory (if the weight was $\frac{1}{L}$, then it would be the average). This first cosine coefficient serves as a baseline, c.f. the intercept in a linear regression. As k increases beyond one, the resulting cosines gradually increase in frequency, $k = 2$ yields a cosine that completes one quarter of its cycle within the signal, $k = 3$, yields a half cycle and so on (see Figure 6). DCT produces $k = L$ components, so the number of cosine components depends on the length of the signal. However, the magnitude of the higher frequency components may be quite small for signals of slow moving articulators.

$$y(k) = w(k) \sum_{n=1} \cos \frac{\pi(2n-1)(k-1)}{2L} \quad k = 1, 2, \ldots, L \tag{1}$$

where

$$w(k) = \begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \leq k \leq L \end{cases} \tag{2}$$

DCT has a known inverse function, iDCT, which can be used to simulate trajectories from DCT components (= Equations (3) and (4)).

$$x(n) = \sum_{n=1} w(k) y(k) \cos \frac{\pi(2n-1)(k-1)}{2L} \quad n = 1, 2, \ldots, L \tag{3}$$

where

$$\begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \leq k \leq L \end{cases} \tag{4}$$

We make use of iDCT to assess how many DCT components are necessary to faithfully represent the actual articulatory trajectories. We do this by fitting DCT components to a set of trajectories and then resynthesizing using iDCT with progressively more DCT components. In this way, we can observe how increasing the number of DCT components improves the precision of the representation. Figure 5 shows representative results, from one speaker and one item ([ʃutokou] produced by Speaker 7). The improvement from 1 DCT component to 2 is substantial, as is the improvement from the 2 components to 3 components. With four components the correlation between the raw trajectories and the iDCT-simulated trajectories reaches $r = 0.99$. In our case, only a small number of DCT components (3 or 4) are required to faithfully represent articulatory trajectories over the target VCVCV window. This result is similar to past studies, which have modelled trajectories of similar duration and linguistic complexity using either 3 (Shaw & Kawahara, 2018a) or 4 (Shaw & Kawahara, 2018b; Kawahara et al., to appear) components.

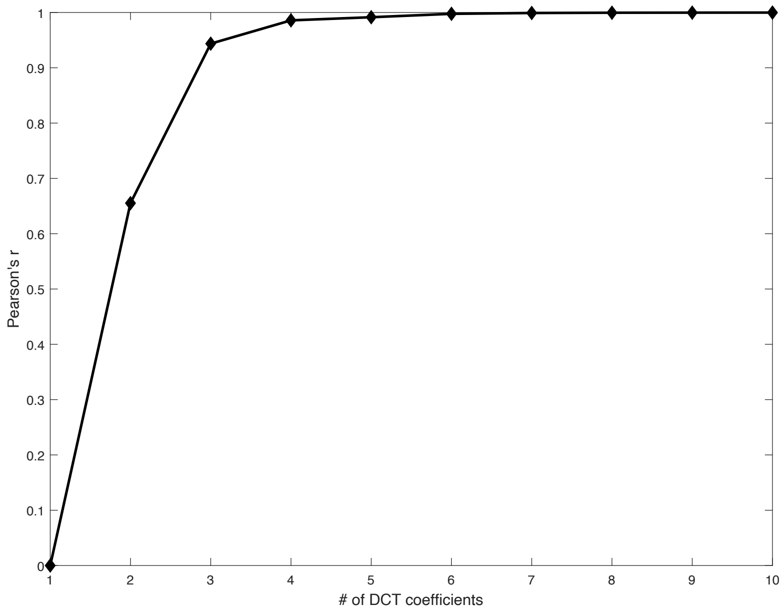

Figure 5: The increase in Pearson coefficients between the number of DCT components and the correlation between actual trajectories and simulated trajectories.

We can also use iDCT to illustrate how each DCT component contributes to the representation of the articulatory trajectory. The top panel of Figure 6 shows the average articulatory trajectories for each item of the dyad, [ʃutokoo] (left) vs. [ʃudooken] (right).

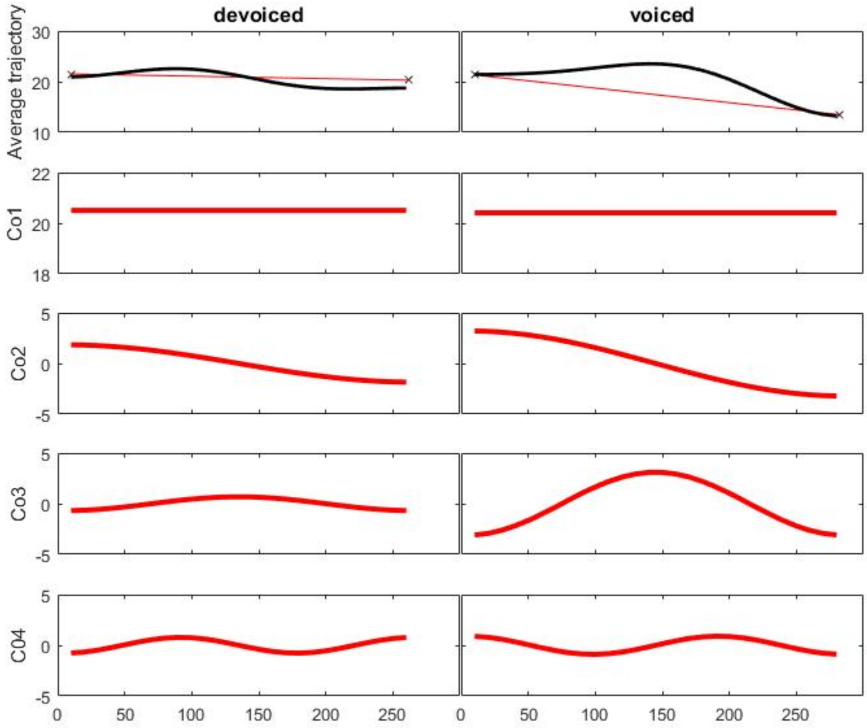

Figure 6: A sample comparison between the four DCT components of articulatory trajectories of devoiced and voiced tokens (averaged). The top panel shows the signal, with the 'x' marking the average height at the begining and end of the trajectories and the line between the 'x's indicating linear interpolation.

Given this dyad, we can observe that the average change in tongue dorsum height over time, shown in the top panel, is noticeably different between devoiced and voiced items. For the voiced item (right), the tongue dorsum rises in the middle of the trajectory for [u]. For the devoiced item, there is less variation in the positional signal over time. For reference, the "x"s in the top panel show the average position at the start and end of the analysis window. The straight line connecting the x-points is equivalent to a linear interpolation of spatial position across the analysis window. The panels below the trajectory show the contribution of each DCT component to spatial modulation of the signal. The duration of the simulated iDCT is based on the average duration of the tokens.

Comparison across devoiced and voiced items reveals similar modulations for the first coefficient (Co1) and the second coefficient (Co2). The main difference is in the third (Co3) and fourth (Co4) coefficients. Co3 picks up on the large rise for [u] in the voiced case.[8] The magnitude of the rise contributed by Co3 is greatly reduced for the devoiced item compared to the voiced item. Finally, the fourth DCT coefficient (Co4) is also quite different between voiced and devoiced items but it has only a small effect on spatial position overall.

The next step is to assess whether the devoiced item contains a vowel target or not. To do this we set up stochastic generators of our competing hypotheses, which we use for Bayesian classification. The "target present" hypothesis is based on the voiced member of each dyad. Specifically, since we have multiple repetitions of each item, we can calculate a distribution over each DCT component. The normal distribution is characterized by a mean value and a standard deviation. Thus, the mean and standard deviation of each DCT component characterizes a normal probability distribution function. For the "target absent" case, we adopt the common assumption that, in the absence of phonological specification, the trajectory will interpolate between surrounding targets (Choi, 1995; Cohn, 1993; Keating, 1988; Pierrehumbert & Beckman, 1988). We therefore construct probability distributions for the "target absent" hypothesis that capture a realistically noisy interpolation. For each token of a devoiced item, we fit DCT components to the straight line connecting the position at the onset and offset of the analysis window.[9] The average of these components defines the probability distributions for the "target absent" hypothesis. The standard deviation for the distributions is computed from the devoiced trajectories in the same manner as for the voiced item. Consequently, the probability distributions that characterize the "target absent" hypothesis are defined by linear interpolation (means of the distribution) and the variability around each DCT component in the data. An example

[8] We note however that it is not necessarily the case that each DCT coefficient has to have a meaningful linguistic interpretation; neither is it the case that we have reasons to believe that Co3 is solely responsible for representing the tongue dorsum raising gesture of [u].

[9] See Pierrehumbert (1980) and Myers (1998) for cases of non-linear interpolation. We will reexamine this analytical choice of ours in §5.3.

of the resulting distributions is provided in Figure 7. The horizontal axis is the value of the coefficient, i.e., y in Equation (1), and the vertical access is probability.

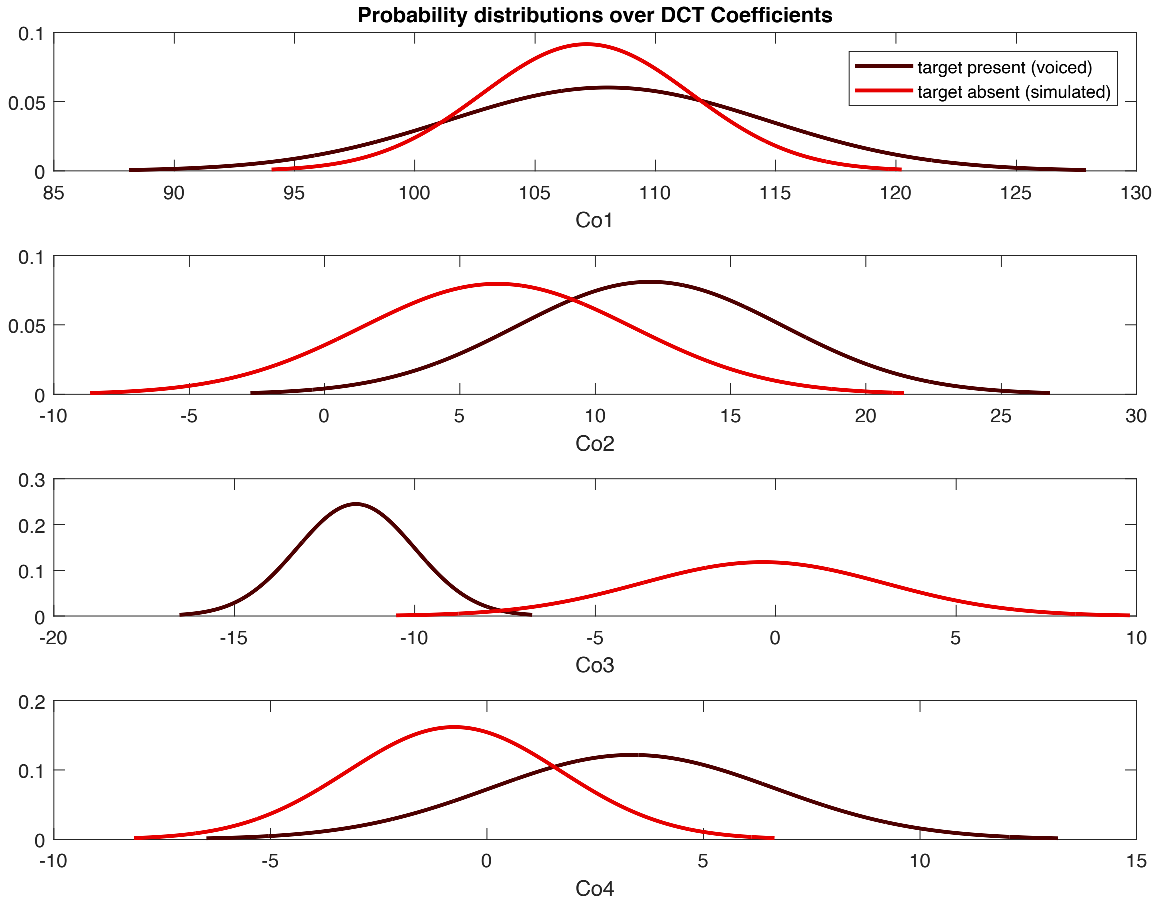

Figure 7: Probability distributions for DCT coefficients for the two competing hypotheses. The "target present" condition is based on the voiced vowels. The "target absent" condition is based on linear interpolation and the level of variability in the devoiced vowels.

We observe that the distributions for Co1 between the two conditions overlap heavily. For Co2, there is a small difference between the "target present" distributions, based on voiced vowels, and "target absent" distribution, based on linear interpolation. The largest difference appears to lie in Co3. Naturally, the mean of the "target absent" distribution is very close to zero, and the same goes for Co4. This is because there is no rise for the straight line fit connecting the positional signal at the onset and offset of the analysis window. The "target absent" Co3 distribution is also more variable than the corresponding "target absent" distribution—this difference reflects greater variability across devoiced tokens than voiced tokens in whether the trajectory showed a rise characteristic of a vowel or not.

As the final step of the computational analysis, for each devoiced token, we determined the posterior probability of a vowel height target, based on Bayesian classification of the tongue dorsum trajectory (=Equation (5)). The posterior probability of the targetless hypothesis given the set of DCT coefficients (the left term of the Equation) is expressed as the prior probability of the targetless hypothesis—always set to be 0.5 in the current analysis, i.e, a uniform prior—multiplied by the product of the conditional probabilities of each DCT coefficient given the targetless hypothesis (i.e. linear interpolation), normalized by the denominator term. The classifier was trained on the distributions described above (see Figure 7) for voiced tokens, which unambiguously contain a vowel target, and a noisy null hypothesis, defined as linear interpolation across the target interval.

$$p(T|Co_1,\ldots,Co_n) = \frac{p(T) \times \prod_{i=1}^{n} p(Co_i|T)}{\prod_{i=1}^{n} p(Co_i)} \tag{5}$$

To summarize, the approach described in this subsection assigns a probability of target absence to each token. It does so by considering the probability that the token follows a linear interpolation as opposed to the trajectory of voiced vowels.

4 Results

Figure 8 shows the posterior probability of target absence for each condition by each speaker. The figures are violin plots which show the distribution of posterior probabilities of target absence. Points around the high y-axis region are tokens with a high probability of target absence, i.e., lingual movements that can be characterized as linear interpolation through the devoiced portion of the signal. Those at the bottom of the y-axis are tokens that have a high probability of a vowel target, i.e., lingual articulations that resemble the voiced tokens. Those in the middle range are intermediate between target present and target absent, indicating a spatially reduced vowel target.

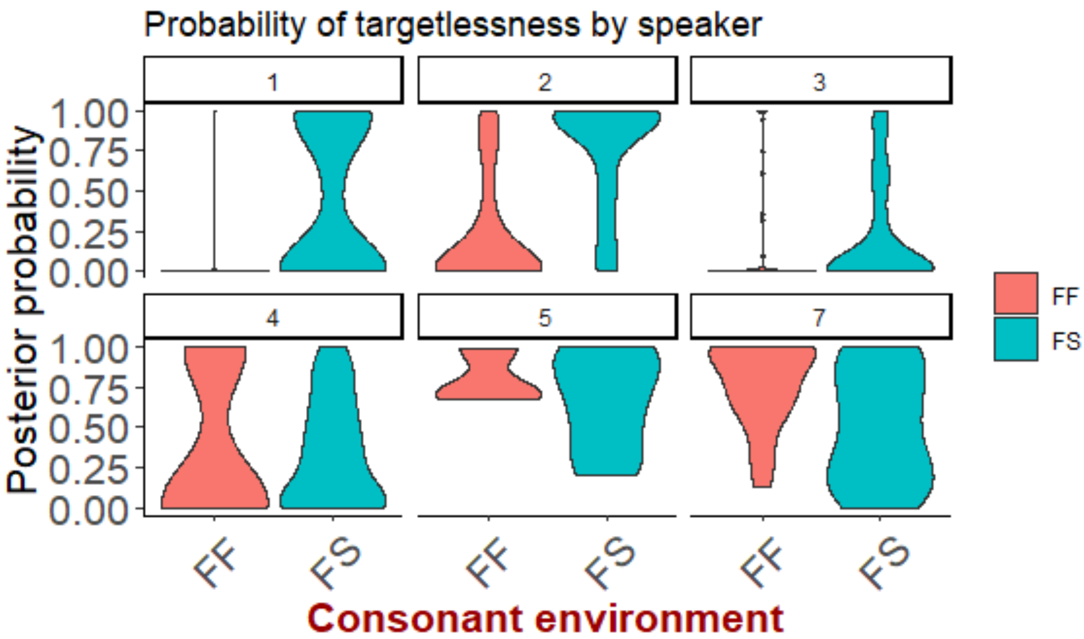

Figure 8: Posterior "target absent" probability for each condition by speaker. FF=Fricative-Fricative; FS=Fricative-Stop.

We observe that, as with Shaw & Kawahara (2018b), the distribution of posterior probabilities is bimodal. Across speakers, there tends to be a large probability mass at the high end of the probability scale (e.g., FS items for speaker 2 and speaker 5), at the low end of the probability scale (e.g., FF items for speaker 2, all items for speaker 3, FS items for speaker 4), or both (e.g., FS items for speaker 1, FF items for speaker 4). In many conditions, items skew towards the high and low ends of the scale. This is not to say that there are no intermediate items, which we take to be reduced. There are several cases with probability mass in the middle range , e.g. the FF condition for Speakers 5 and 7. Overall, however, the by-speaker view shows a tendency to either fully retain the lingual gesture or entirely lose it. The one possible exception is FF items for speaker 5, the only plot of 12 in Figure 8 which does not have the majority of the probability mass at one end of the scale. This result replicates the findings by Shaw & Kawahara (2018b) with a new set of speakers and an expanded set of stimuli. Recall that the study by Shaw & Kawahara (2018b) examined only four dyads; the current results are based on twelve dyads.

How the flanking consonants influenced targetless probability varied between speakers. Speaker 1 showed almost no targetless tokens in the FF condition, but showed some targetless tokens in the FS condition. This pattern—more targetlessness in the FS condition than in the FF condition—accords well with the prediction laid out in §1.3. Speaker 2 shows a similar, and perhaps clearer, pattern; this speaker showed rather consistent target-present production in the FF condition, but typically deleted tongue dorsum raising gesture in the FS condition. The pattern exhibited by Speaker 3 is less clear, but is also consistent with the hypothesis presented in §1.3: almost no targetless tokens in the FF condition, but greater probability of targetlessness in the FS condition. These three speakers thus confirmed the hypothesis that we formulated in §1.3.

However, not all speakers behaved as we hypothesized. Speaker 5, especially in the FF condition, seems to show some tokens whose posterior probabilities are in the middle range—those tokens that are neither clearly targetless nor have a full target. Speakers 4 and 7, especially the latter, showed a pattern that is opposite from what is predicted from the considerations discussed in §1.3—more targetless tokens in the FF condition than in the FS condition. Thus, looking across the six speakers, we observe speaker-specific variation in whether FF or FS environments conditions more deletion of the tongue dorsum raising gesture.

Figure 9 shows the results by item. From this plot we can see some variability across items as well. For example, [Fusagaru], the only verb in the item list, shows the lowest probability of targetlessness. Many words show fairly sharp bi-modal patterns, with some tokens showing high probability of targetlessness and others showing high probability of full targets with few intermediate tokens. This bi-modal pattern applies especially clearly to [Futa], [Futan], [Futon], [ʃusa], and [ʃutokou]. In contrast, most tokens of [ʃusai] are intermediate, with few extreme probabilities in either direction.

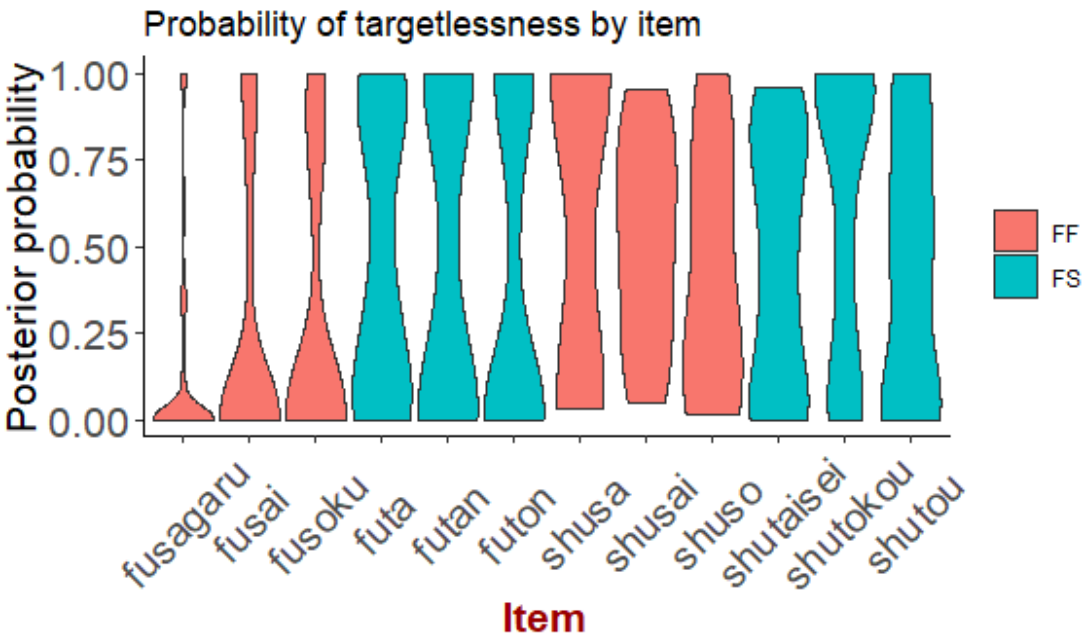

Figure 9: Posterior "target absent" probability by item. "f" and "sh" are used in the figure in place of [F] and [ʂ], respectively.

To assess the overall results statistically, we fit a series of nested linear mixed effects models in (6). The results of model comparisons appear in Table 2. The baseline model, *m*0, was compared to *m*1; then *m*2 and *m*3, which have the same number of parameters, were compared to *m*1. Finally, *m*4 was compared to *m*3. The dependent variable was the posterior probability of deletion. Since probabilities are bounded dependent variables (upper bound of 1; lower bound of 0), we also ran the same models on arcsin-transformed probabilities. The same pattern of results came out of both raw and transformed probabilities. For reasons of space we report results based on the non-transformed probabilities. The key fixed effect of interest was the consonant environment, coded as a two level factors, FF vs. FS ("*Cond*"). Speakers and items were treated as random intercepts.

$$
\begin{aligned}
m0 &: post ¾ (1 j speaker) + (1 j item) \\
m1 &: post ¾ (1 + Cond j speaker) + (1 j item) \\
m2 &: post ¾ Cond + (1 + Cond j speaker) + (1 j item) \\
m3 &: post ¾ C1 + (1 + Cond j speaker) + (1 j item) \\
m4 &: post ¾ C1 Ł Cond + (1 + Cond j speaker) + (1 j item)
\end{aligned}
\quad (6)
$$

Table 2: Summary of model comparisons.

	df	AIC	BIC	logLik	deviance	^{2}	^{2} df	p
$m0$	4	464.7	481.7	-228.3	456.7	–	–	–
$m1$	6	402.6	428.1	-195.3	390.6	66.07	2	$< :001$
$m2$	7	404.1	433.8	-195.0	390.1	0.53	1	$n:s:$
$m3$	7	400.4	430.1	-193.2	386.4	4.25	1	$< :05$
$m4$	9	403.7	441.9	-192.8	385.7	4.95	3	$n:s:$

The baseline model, $m0$, includes only the random effects. The next model, $m1$, adds a by-speaker random slope for the fixed effect, i.e. surrounding consonants (FF vs. FS) to this model. The by-speaker random slope improved the model significantly. This result indicates that speakers show different sensitivities to the consonantal environments. As we observed in Figure 8, some speakers (e.g. Speakers 1 and 2) show less deletion in FF than FS environments, while others (Speakers 4 and 7) show the opposite pattern.

Because the effect of consonant environment differs by speaker, the average effect of consonantal environment is not predictive. These statistical comparisons support what we observed in Figure 8: different speakers are sensitive to consonantal environment in different ways.

We also ran models that included the C_1 type ([F] vs. [S]) and the interaction between C_1 and consonant environment ("$Cond$") as fixed factors. The addition of C_1 led to improvement over $m1$, and was marginally significant within the model ($b = 0:098; t = 2:136; p = 0:055$), indicating that deletion probability is slightly higher when C_1 is [S] than when C1 is [F]. The interaction between C_1 and consonant environment ("*Cond*") did not lead to further improvement, indicating that the effect of C_1 is not dependent on the consonant sequence. Thus, our best fitting model, m3, includes a ("*Cond*") as random effect but not as a fixed effect.

Figure 10 shows the by-speaker random slopes for our best fitting model. The x-axis shows the estimate for FS sequences. As we observed in the violin plots of probabilities (=Figure 8), Speakers 1 and 2 have positive estimates, indicating that deletion is more likely in FS sequences than in FF sequences. Moreover, the confidence intervals around the estimate do not overlap with zero. Additionally, as we also observed above, Speaker 7 shows the opposite pattern. This speaker has a negative estimate, which also does not overlap with zero, indicating significantly higher probability of targetlessness in FF sequences than in FS sequences. The other speakers have estimates that overlap with zero, indicating an effect that is not statistically significant.

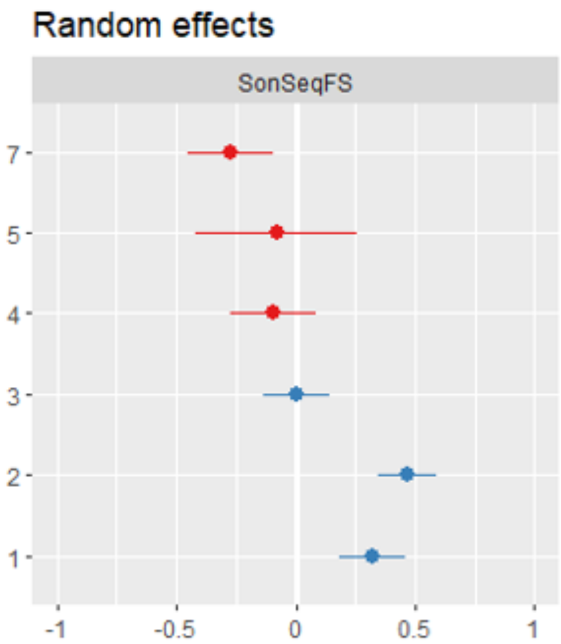

Figure 10: By-speaker random slopes for the effect of sonority sequencing (=*Cond*). The estimate is for the FS condition, relative to FF.

In summary, consonant environment had a significant impact on deletion probability, but the direction of the effect was not uniform across speakers. Some speakers showed consistently more deletion in FS, as predicted, others showed more deletion in FF, or no effect of consonant context.

5 Discussion

5.1 Summary

The current experiment replicated the core finding of Shaw & Kawahara (2018b) with a new set of speakers and an extended set of stimuli. The posterior probability of vowel presence/absence showed a bimodal distribution for many speakers (see, Figure 8) and items (see, Figure 9). One mode was centered on the low end, near zero probability of vowel absence. These devoiced vowel tokens were produced with tongue height trajectories very similar to voiced vowels. The other mode of the distribution was centered on the high end, indicating that the tongue height trajectory resembled our noisy null hypothesis, a linear interpolation between flanking vowel targets. These modes of the posterior probability distribution represent endpoints on a continuum from a full target to no detectable vowel target. A mono-modal distribution centered between 0 and 1 would have provided evidence for consistent vowel reduction, i.e., a vowel height target of reduced magnitude. Although we did also see some tokens with intermediate probabilities, the variation clustered more around the high and low ends of the scale, a similar pattern reported in Shaw & Kawahara (2018b).

The results also revealed some systematic patterns in how the flanking consonants influence deletion probability. The design of the study featured conditions contrasting devoiced vowels intervening between fricative-fricative (FF) sequences and fricative-stop (FS) sequences. The original hypothesis developed in §1.3 is that was that we would observe more deletion in FS sequences than in FF sequences. Recall that, to the extent that we can conceive of deletion as an extreme instantiation of devoicing, either in terms of reduction or enhancement, we would expect targetless tokens to be more likely in the FS condition than in the FF condition, because devoicing is more likely in this environment.

Syllable contact laws (Murray, 1988; Murray & Vennemann, 1983), if Japanese speakers are sensitive to them, also predict this pattern. Our hypothesis was also motivated by an empirical observation. Shaw & Kawahara (2018b) found that, even though the speakers in the study differed substantially in their individual rates of vowel deletion, all speakers deleted devoiced vowels more often in [ʃutaisei], resulting in a FS consonant sequence, than in [Fusoku], resulting in a FF sequence.

The current study revealed inter-speaker variability with respect to the prediction laid out in §1.3: some speakers showed more targetless tokens in the FS condition than in the FF condition (Speakers 1 and 2), as we initially hypothesized, and some speakers showed the opposite pattern (Speaker 7, and to a less clear extent, Speaker 4).

Our items in the FF and FS condition both featured two fricatives, [F] and [ʃ]. Although we did not predict this differences, there was a significant effect of fricative, with higher deletion probability following [ʃ] than [F]. Moreover, this effect is significant in a group analysis while consonant sequence was only significant as a by-subject random slope. Quite possibly, the observed difference in deletion probability between [ʃutaisei] and [Fusoku] in past work as well is attributable not to the consonant manner sequence, FF vs, FS, but to the identity of the initial consonant.

5.2 Time and target undershoot in DCT representations

Our approach to analyzing time-varying kinematic data in terms of discrete hypotheses makes use of a low parameter stochastic representational space. Time varying signals, in this case tongue dorsum height trajectories, are represented as modulations of frequency components, using DCT. The DCT coefficients effectively represent the signal with high precision but without directly encoding the temporal duration of the trajectories. Instead, time is indirectly encoded in the frequencies of the DCT components. The representation of time is indirect because it comes in the form of what frequencies are represented in each component, which is dependent on the analysis window.

We represented all trajectories in this study using just four DCT components. Since the frequency of the DCT components vary as a function of the length (in samples) of a trajectory (see (1)), they have the potential to indirectly encode the duration of the trajectory. For example, past work has shown that DCT representations alleviate the need to represent temporal duration independently. For example, Watson & Harrington (1999) compared several methods of representing time-varying formants, including DCT representations, in a study of Australian vowels. They showed that adding vowel duration to the representation of Australian vowels improved machine classification in many cases. When Australian vowels were represented by measurements of formants at percentages of total vowel duration, vowel duration was needed as an additional factor to reach a high-level of classification accuracy. This is because several Australian vowel pairs have very similar (possibly indistinguishable) vowel quality but differ in duration (Cox & Fletcher, 2017). However, when Watson & Harrington (1999) represented the same vowels with DCT components only, vowel duration did not improve classification accuracy. Two DCT components fit to the first and second formants were sufficient to classify all 19 Australian vowels, including vowels differentiated primarily by duration.

Since DCTs can represent both the spatial modulation and the temporal duration of a signal, we cannot know if one of these dimensions or the other had a dominating influence on our classification results. Although high vowel devoicing in Tokyo Japanese occurs at both fast and slow speech rates (Fujimoto, 2015), we do not know if vowel deletion is likewise rate independent. Conceivably, the probability of detecting a vowel movement decreases at fast rates due to target undershoot (Lindblom, 1963; Moon & Lindblom, 1994). To investigate this, we evaluated the correlation between the duration of our target intervals, as a measure of local speech rate, and the posterior probability of deletion. Figure 11 shows a scatter plot of these two variables. There was a weak negative correlation ($r = 0.11; p < .05$), indicating that the probability of targetlessness decreases at slower speech rates (longer duration).

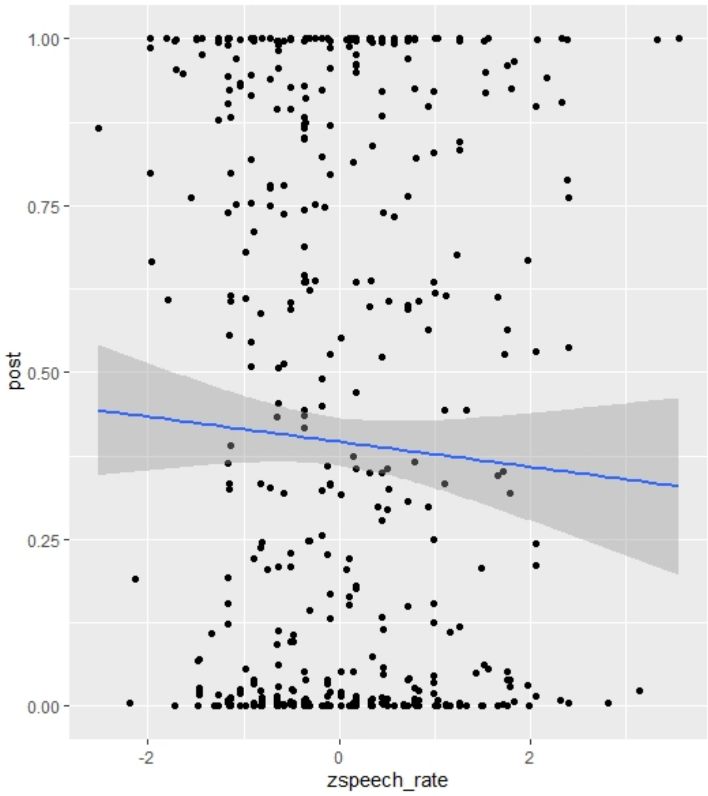

Figure 11: Correlation between speech rate, represented by a Z-scored of target trajectory duration (x-axis) and the posterior probability of targetless (y-axis).

To further investigate the influence that speech rate might have on our deletion probability results, we subsetted the data into relatively short and relatively long tokens. Our short-ish tokens were those that were less than one standard deviation from the mean token duration; our long-ish tokens were those that were greater than one standard deviation from the mean. This subsetting procedure produced 74 tokens (14.4% of the data) for the short group and 76 tokens (14.8%) for the long group. We looked at the distribution of long and short tokens across speakers and found that all speakers produced some tokens that fell into the long group and some that fell into the short group. The mean duration of the CV interval in the short group was 228 ms. The mean duration of the CV interval in the long group was 362 ms. Figure 12 compares the posterior probability of deletion for the long (slow local speech rate) and short (fast local speech rate) data subsets. Consistent with the weak correlation between speech rate and targetlessness across the entire corpus, we see a slight increase in targetlessness probability for the short data subset. This is the case for both FF and FS consonant manner sequences. Notably, however, a substantial number of tokens still show a high probability of targetlessness at slow speech rates. This indicates that while increased speech rate may contribute to targetlessness, based on the diagnostic methods employed here, there are still tokens that approximate a linear interpolation trajectory even at the slowest speech rates in the data set. This indicates that, like high vowel devoicing, vowel deletion, or at least extreme reduction of the tongue dorsum height target, also occurs that slow end of natural speech rate variation. This result implies that whether or not to retain a tongue dorsum gestures is under speakers' control, rather than an automatic consequence of fast speech.[10]

5.3 Minimal paths for targetless trajectories

One of the challenges of assessing whether the tongue dorsum height target is completely absent or just heavily reduced is that there are no unequivocal FF or FS sequences in Japanese that could serve as a baseline for assessing whether pronunciation of /FuF/ and /FuS/ deviate enough from these underlying forms to conclude that they are indeed [FF] and [FS]. Our approach to this challenge is to simulate tongue dorsum trajectories that interpolate between vowels, V1 and V2, in /V1CCV2/. Our simulations in this paper are based on two assumptions: (1) first, movements take the minimal path between targets and (2) second, like all biological signals, there will be variability in the movement trajectory. We calculated the minimal path as a linear interpolation between vowel targets and we modelled variability as random deviations from the minimal path. The magnitude and structure of the random deviations are based on the devoiced tokens in our corpus. In this way, the variability injected into our simulations has the same item-specific and speaker-specific properties of our corpus. The difference between the vowel-absent class, as we simulated it, and the devoiced tokens in our corpus, is that the tongue-dorusm trajectory in

10

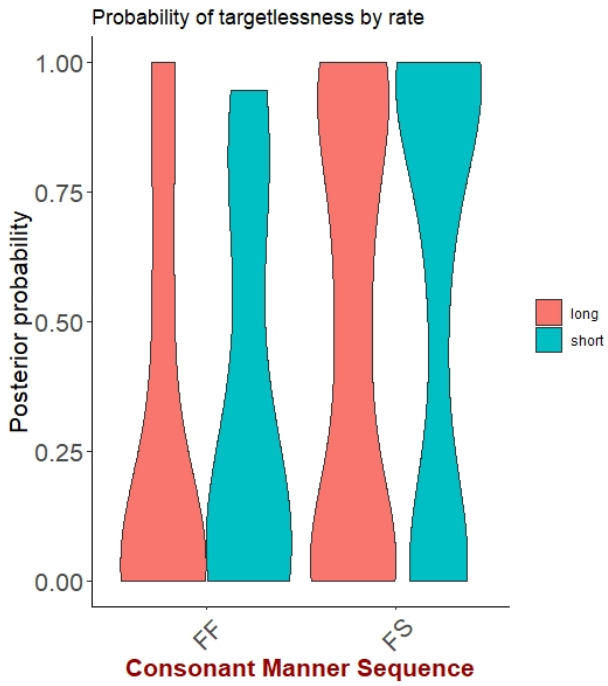

Figure 12: Posterior probablities of the short-ish subset and long-ish subset.

the vowel-absent class is always guided by the minimal distance between V1 and V2. The degree to which the actual tongue dorsum trajectories in our devoiced tokens also follow a realistically noisy actuation of the the minimal distance path or whether they instead move towards an elevated tongue dorsum height target for [u] is represented in the results of our Bayesian classification. A substantial number of tokens were classified as belonging to the minimal distance path.

Our decision to simulate the vowel-absent tongue dorsum trajectory as taking the path of minimal distance between flanking targets is intended to be a theory-neutral decision. It is also possible to apply our method of analysis by simulation and classification with different theoretical assumptions about what the vowel-absent trajectory should look like. Here, we consider the predictions of Task Dynamics (Saltzman & Munhall, 1989) as implemented in the Task Dynamics Application (TADA: Nam et al. 2004, 2012) . One property of this model is that articulators that are not under direct phonological control (i.e, by a gesture, in the sense of Articulatory Phonology: Browman & Goldstein 1986 et seq.) at a particular time are driven to a rest position by a neutral attractor. Because of the neutral attractor, there are conditions under which articulators will not necessarily follow the minimal path between targets. Instead, articulators will return to a neutral position until they are brought under control by another gesture. To explore how TADA predictions for the vowel-absent case might differ from linear interpolation for the items in our study, we ran a series of TADA simulations.

The first TADA simulation compares [eFta] and [eFuda]. There are a number of manipulable parameters in TADA, and variation in some of these parameter settings has been hypothesized to capture cross-language variation, i.e., language-specific phonetics (Iskarous et al., 2012). To minimize researcher degrees of freedom (Roettger, 2019), we used default TADA gestural parameters whenever reasonable for Japanese. For the [eFta] vs. [eFuda] comparison, we used default parameters for [e], [f] for [F], [t], [d], and [a]. The only gesture that required manipulation to approximate Japanese-specific phonetics was [u]. The default [u] in TADA produces a much longer vowel, 300 ms, than is typical in Japanese, and it produces a vowel with lip protrusion. To adapt the gesture parameter settings for Japanese [u], which is much shorter, ca. 50 ms (e.g. Shaw & Kawahara 2019), and lacks lip protrusion (e.g. Vance 2008), we eliminated the lip protrusion gesture and shortened the activation duration of the tongue body gesture. The gesture parameter values for all simulations are provided in the supplementary materials.

Figure 13 compares the trajectories for [eFta] and [eFuda] simulated by TADA. The top panel shows the simulated waveform. The bottom three panels show kinematic trajectories in the vertical dimension for the tongue dorsum, tongue tip and lower lip. The tongue dorsum trajectory for [eFta] has a mid-level plateau for [e], in the temporal window from 0 to 250 ms, and then falls to [a]. The tongue dorsum trajectory for [eFuda] starts with a similar plateau for [e] but then rises for [u]. The peak of the rise comes near the end of the voicing period for the vowel and remains rather high during the [d] before falling for [a]. The data simulated with TADA are qualitatively quite similar to our experimental data. For comparison with representative tokens from the experimental data, see Figure 3. For this particular case, our theory-neutral choice of linear interpolation for "vowel-absent" tokens is quite similar to the TADA simulations, which also show a roughly linear trajectory. It should be noted, however, that this linearity is not a general prediction of TADA. It follows in part from the properties of our stimulus items. The progression of vowel height targets from mid, [e], to low, [a], does not involve a neutral attractor driving the tongue dorsum height away from the minimal path between these vowels. For items such as [eFta], there would be little difference between using linear interpolation between flanking vowels and using TADA simulations, with default gesture parameters.

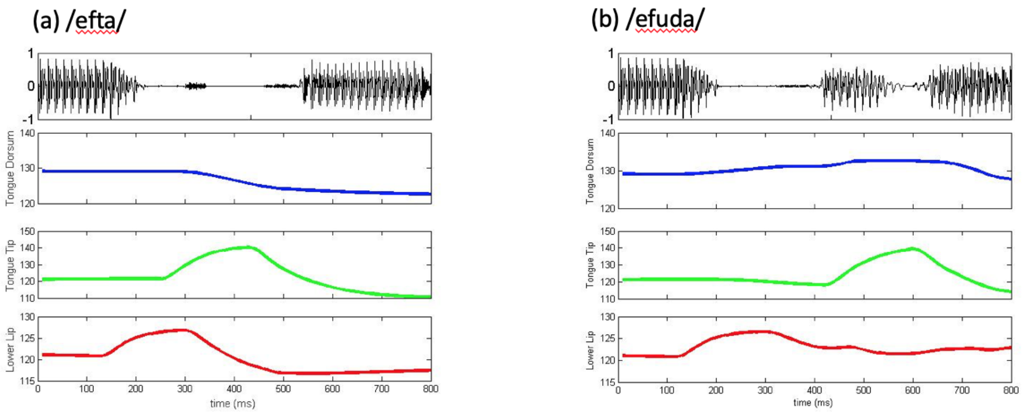

Figure 13: TADA simulations of [eFta] and [eFuda]

We now move on to [eSta] and [eSuda]. Figure 14 shows TADA simulations of these items. The top two panels show simulation results with default gesture parameters for all segments except for [u], which used the same Japanese-specific parameters described above. Of relevance is that the default gestures for [S] include both a tongue body gesture and a tongue tip gesture. For Japanese, our materials were not designed to assess the presence/absence of a tongue body gesture for the fricative, [S], directly (see *S*5.5 for an indirect attempt). The Japanese fricative has different acoustic and articulatory properties from the English post-alveolar fricative, but it is unclear whether the difference is due to the tongue body gesture or to other aspects of fricative production, including a labial component, tongue-tip constriction area, or relative degree of tongue grooving. Because of this uncertainty, we also ran TADA simulations with the fricative unspecified for a tongue body gesture. This result is shown in the bottom panel of Figure 14

When [S] was simulated without a tongue body gesture, the difference in tongue dorsum trajectories between [eSta] and [eSuda] is nearly identical to the difference found for [eFta] and [eFuda]. That is, the tongue dorsum height trajectory follows a roughly linear path from [e] to [a] in [eSta] but it rises for [eSuda] . However, when [S] is specified with a tongue body gesture, then we see a rise in the tongue dorsum height trajectory in [eSta], which disrupts the linearity of the transition from [e] to [a], even in the absence of [u].

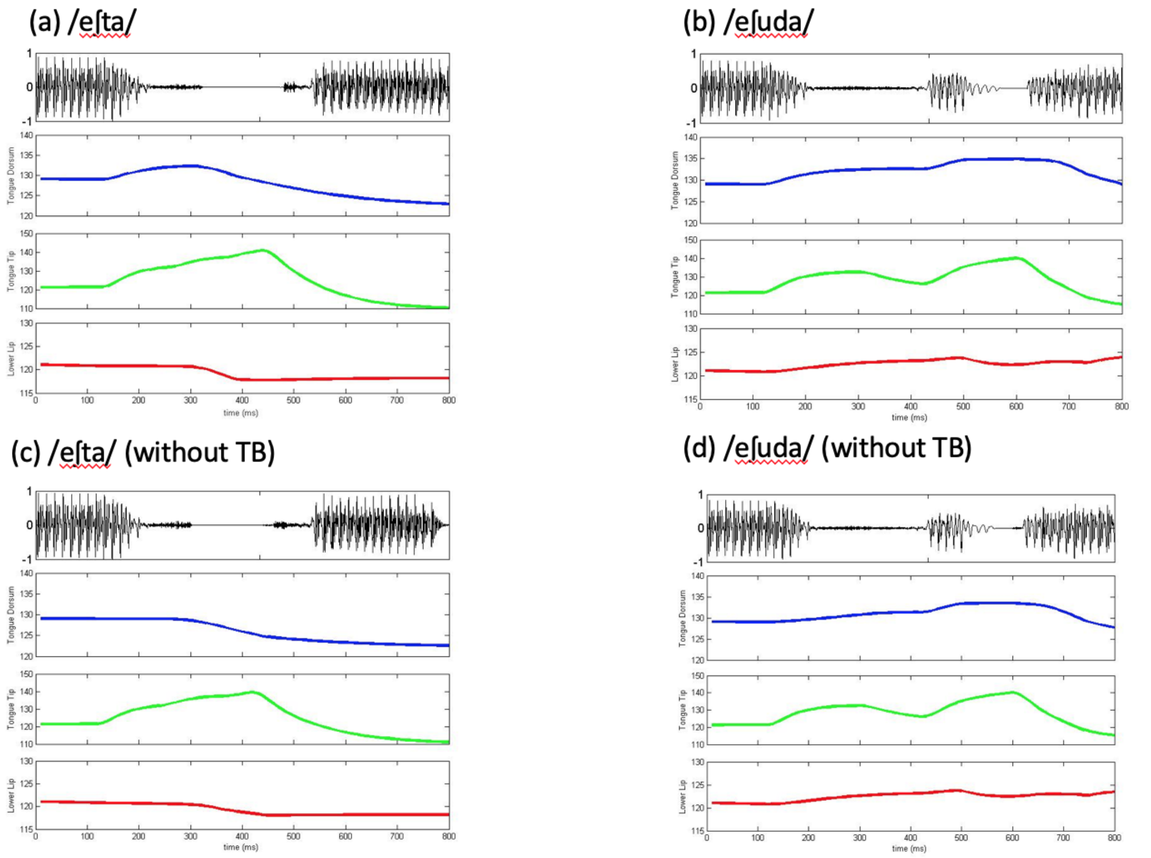

Figure 14: TADA simulations of [eSta] and [eSuda] with (top row) and without (bottom row) TB gesture.

The case of [S] specified with a tongue body gesture allows us to consider how using a theory-specific alternative to the minimal path assumption might influence our results. If we used (a stochastic version of) the TADA simulation trajectory for [eSta] and [eFta] as the basis for our Bayesian classification (instead of linear interpolation), we would introduce a bias in deletion likelihood towards the [S] environment over the [F] environment. This is because, to detect a vowel in the [F] environment, the trajectory would only have to rise above the linear trajectory in the TADA simulation (Figure 13, [eFta] panel). However, to detect a vowel in the [S] environment, the trajectory would have to rise above not just the linear trajectory between vowels but also above the magnitude of the tongue body gesture for [S]. Deviations from minimal path would still be classified as deletion, if the magnitude of the deviation did not exceed the tongue body magnitude for [S]. In contrast, relative to using a TADA baseline, if there actually is a tongue body gesture for [S], the minimal path method is biased towards finding more vowel deletion in the [F] environment than in the [S] environment. This is because increases in tongue body height, including those due to [S], will count as deviation from the minimal path, and push classification towards the vowel present case.

Using the minimal path method, we observed significantly greater deletion in the [S] environment than in the [F] environment. If we had used a TADA-baseline with a tongue body gesture for [S], this result would probably have been even stronger. On the other hand, if we had used a TADA baseline without a tongue body gesture for [S], then there is really not much difference between the minimal path method and a TADA baseline. However, we reiterate that the similarity between TADA and minimal path is not a general result—it is particular to the materials that we selected for this experiment. Additionally, the above conclusions are based on default gesture parameters (with the exception of [u]), which are appropriate for English, but might require fine-tuning in order to capture systematic differences across languages. Generally, there may be conditions under which a minimal path baseline is inappropriate, or, at least, is inconsistent with the Task Dynamics framework, as implemented in TADA.

With the above caveats in place, we conclude that the finding of more deletion in the [S] environment than in the [F] environment is likely robust to variation in how we might simulate the vowel absent scenario. If there is indeed a tongue body gesture for [S], the minimal path method is biased against our finding, and yet it still emerged as statistically significant.

5.4 Tongue dorsum trajectories for voiced vowels

In the last sub-section, we discussed how we simulated, for the purpose of classification, trajectories lacking a vowel target. The other relevant factor in classifying devoiced trajectories using our method is the trajectory of the corresponding voiced vowel. We defined a separate classifier for each combination of speaker and item. This allows us to incorporate any speaker-specific variation into the analysis. How a particular devoiced trajectory is classified depends both on the degree to which it deviates from the minimal path as well as the degree to which it deviates from the corresponding voiced vowel. Correspondence in this case is based on the materials—we selected near minimal pairs matched on as many relevant properties as possible. To facilitate appropriate generalization of our approach to new data, we discuss some possible non-obvious implications of using a local (by speaker, by item) baseline.

To illustrate the importance of the local baseline, we zoom in on a small subset of the data, just the [F] environment tokens produced by Speaker 2. Recall that Speaker 2 was one of the speakers that produced a particularly sharp bimodal distribution in vowel deletion probabilities and showed the predicted effect of consonant sequence (see Figure 8). Figure 15 shows three panels summarizing tongue dorsum trajectories for Speaker 2. The first panel shows the average tongue dorsum trajecory for voiced and devoiced tokens. This was generated by fitting an SSANOVA, using the GSS package in R (Gu 2002), to the first 150 ms of each token. We choose 150 ms because it is the length of the smallest analysis window for this speaker. The SSANOVA plot shows that, on average, the devoiced trajectories are flatter than for the voiced trajectories. Note that this was not the pattern for all speakers; Speaker 3, for example, showed very little difference between voiced and devoiced trajectories. The second panel breaks down the devoiced tokens by item. Looking across items, we see that [Fusagaru] seems to have the flattest trajectory. From this figure, we might erroneously suspect that [Fusagaru] has the highest probability of deletion. The third panel shows that this is absolutely not the case. In fact, for this speaker, [Fusagaru] has the lowest posterior probability of deletion of any [F]-tokens. This might seem puzzling. Why does [Fusagaru] have a low probability of deletion, given its relatively linear trajectory?

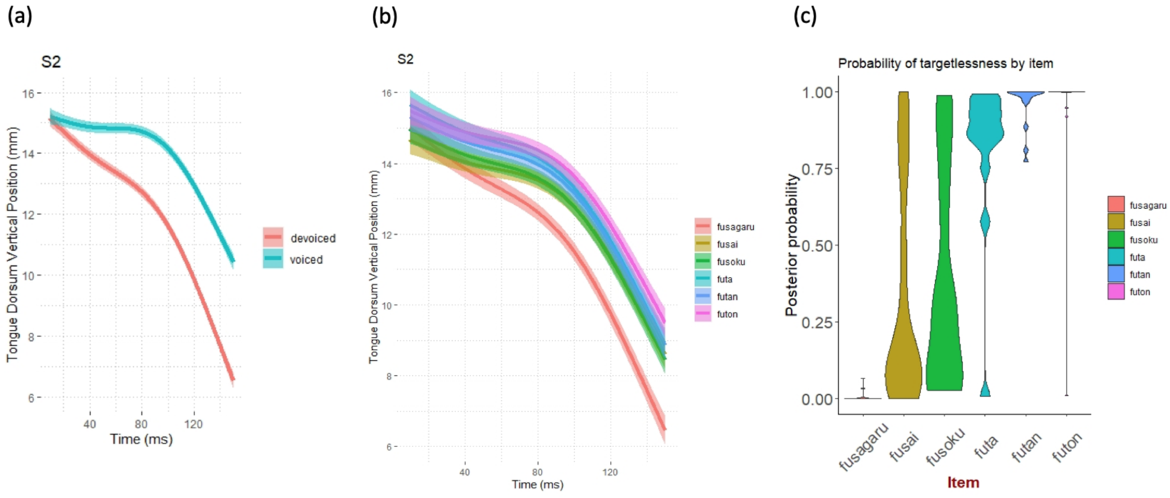

Figure 15: [F]-tokens for S2: (a) shows the average tongue dorsum height trajectory for voiced and devoiced vowels; (b) breaks down the devoiced trajectories by item; (c) shows posterior probability of vowel deletion by item.

The answer is in the patterning of the voiced vowel counterpart for the devoiced tokens. Figure 16 plots [Fusagaru] along with its voiced vowel counterpart [Fuzakeru]. The key observation is that the trajectory for [Fuzakeru], the voiced vowel counterpart to [Fusagaru] in our materials, also has a relatively flat trajectory. Because of this relatively flat baseline for the voiced vowel, the trajectory for [Fusagaru] does not have to depart very far from linearity to be classified as a vowel. The Speaker 2 voiced vowel baseline for [Futa] is quite different. As show in the right side of Figure 16, the tongue dorsum rises substantially for [Fuda], which serves as the voiced vowel baseline for assessing targetlessness in [Futa]. Given this baseline, a [Futa] token that shows only a minimal departure from linearity will still have a higher probability of linearity than of a full vowel.

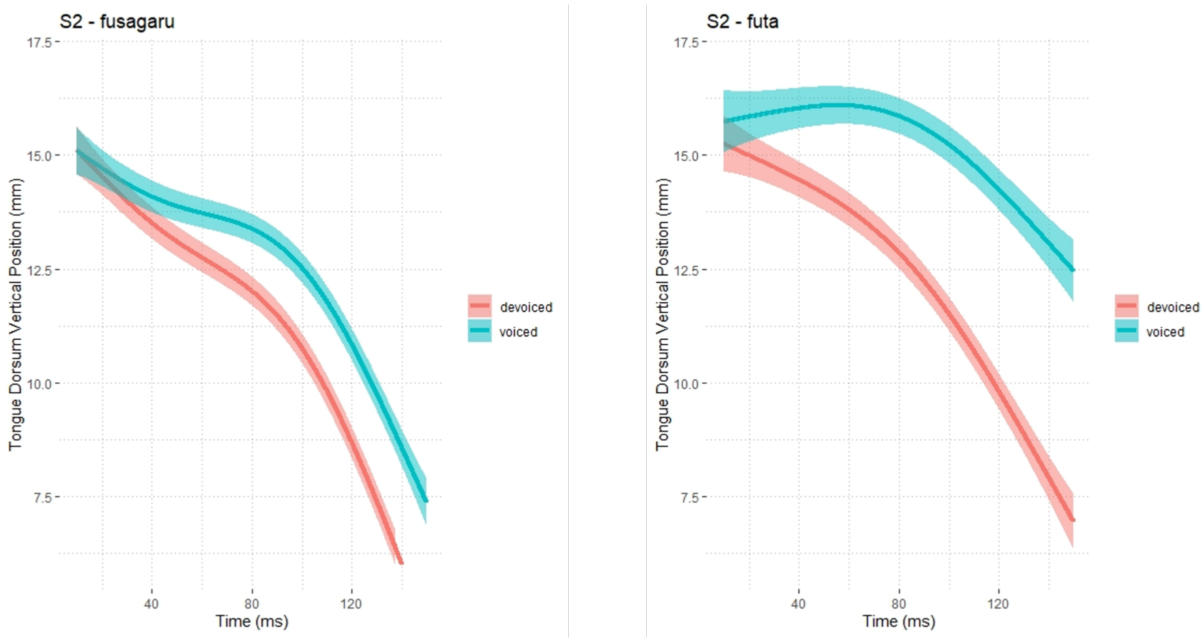

Figure 16: S02 tongue dorsum trajectories for two dyads: the left panel shows [Fusagaru] (devoiced) paired with [Fuzakeru] (voiced); the right panel shows [Futa] (devoiced) paired with [Fuda] (voiced) .

The case above serves to illustrate the role of the speaker- and item- specific baseline in our analytical approach. In assessing whether a given speaker produces a vowel, we pursue a very targeted machine learning approach that factors speaker-specific productions of baseline words in the analysis.

5.5 The effect of fricative place

We now return to the effect that fricative place of articulation had on vowel deletion probability. For starters, we explore an indirect test of whether [ʃ] in Japanese has a tongue body gesture. As illustrated through TADA simulations (Figure 14), whether [ʃ] in Japanese has a tongue body gesture or not is an important consideration in interpreting our results. When we simulated [ʃ] without a tongue body gesture, then the tongue dorsum height trajectory for [eFta] and [eSta] was very similar. As an indirect test of whether Japanese [S] has a tongue body gesture, we compare the distribution of DCT coefficients for all voiced vowel tokens in our corpus. This includes all of the words with voiced vowels that served as item-specific baselines for the devoiced items in both [F]] and [S] environments. Figure 17 compares the distributions. The distributions of all four DCT components are heavily overlapped. Independent t-tests (Welch's two sample) show that differences are not significant for the first three DCT components: 1 ($t = 1.11; p = 0.267$), 2 ($t = 0.406; p = 0.685$), 3 ($t = 1.25; p = 0.214$). Only the fourth DCT component, which explains only a small amount of variance in the trajectories (Figure 5), showed a significant difference ($t = 4.87; p < .001$) across [F] and [S]. Although this result cannot be taken as conclusive evidence for the presence or absence of a tongue body gesture, it does indicate that the trajectories, as represented by DCT coefficients in our classification process, were quite similar across [F] and [S]. This is despite the fact that [F] and [S] tokens were not completely balanced for vowel sequences and other properties (e.g. word length, pitch accent placement, and vowel sequence).

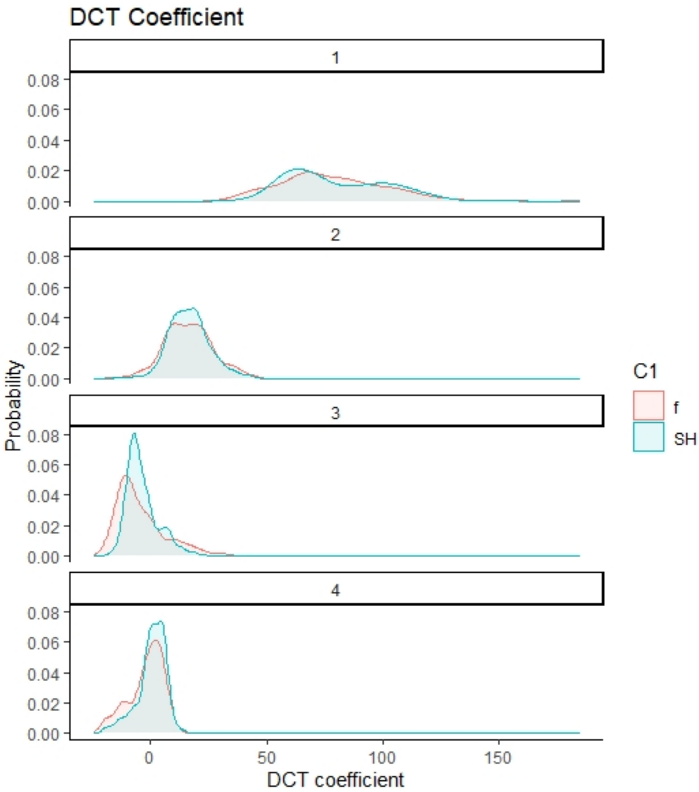

Figure 17: DCT distributions

Given the similarity of the DCT distribution of voiced vowel items across [S] and [F], the difference between [S]-initial items and [F]-initial items can be attributed to the tongue dorsum trajectory in the voiceless items. The trajectory of devoiced vowels is more likely to resemble a linear trajectory between flanking vowels when preceded by [S] than when preceded by [F]. This result is independent of consonant sequence, i.e., FF vs. FS.

One possible explanation for the effect of fricative place on vowel deletion relates directly to the goal of achieving vowel devoicing. While vowel devoicing does not serve a contrastive function, it does serve as a sociolinguistic marker of prestige in Tokyo Japanese (Imai, 2004), and there is evidence that it is under direct control, c.f., devoicing as a passive consequence of overlapping laryngeal gestures, as it may be in some cases of vowel devoicing in other languages (see Fujimoto 2015 and other references cited in introduction). One piece of support for the conclusion the devoicing is actively controlled in Japanese comes from the observation of laryngeal gestures associated with voiceless stops (Fujimoto, 2015). When voiceless stops in Japanese precede voiced vowels, the peak opening of the laryngeal gesture is timed to occur around the release of the supralaryngeal constriction, resulting in long-lag VOT. When a voiceless stops precedes a devoiced vowel, in contrast, the laryngeal gesture of the voiceless stop temporally aligns with the vowel midpoint and increases in magnitude substantially. In devoiced vowels, the laryngeal abduction is greater than two times the magnitude of a voiceless stop preceeding a voiced vowel. The shift in the timing and magnitude of the laryngeal gesture indicates a gesture reorganization that facilitates devoicing.

In contrast to voiceless stops, which show substantial temporal variation between laryngeal and supra-laryngeal gestures, both in Japanese and in the world's languages, the laryngeal and supra-laryngeal gestures of fricatives cannot be temporally displaced so easily. This has consequences for the kinematics of devoicing. In fricative environments, devoicing is not achieved by adjusting the timing or magnitude of the glottal opening, at least not in Tokyo Japanese. Instead, the timing and magnitude of the laryngeal gestures for fricatives is similar when preceding both voiced and devoiced vowels (Fujimoto, 2015). This means that devoicing following fricatives is achieved in some other way.

As an acoustic description, high vowel devoicing following fricative environments can be characterized as a prolonging of the aperiodic energy of a fricative so that it extends across the lingual articulation of the vowel. Articulatorily, maintenance of turbulent airflow is facilitated by narrowing the vocal tract. A key difference between [**S**] and [F] is vocal tract width, i.e. cross-sectional area. Since [**S**] has a constriction in the vocal tract, it naturally conditions a narrow channel that facilitates prolonged turbulent airflow. This is not the case for [F]. Since there is no oral constriction, it is naturally more difficult to sustain turbulent airflow. Specifically, the amount of airflow needed to generate turbulence is a function of the width of the channel, so narrowing the channel means that turbulence can be achieved with less airflow.

Raising the tongue dorsum for [u] narrows the vocal tract and therefore facilitates devoicing, when devoicing is generated from the prolongation of aperiodic energy. Such facilitation is likely more helpful in the environment following [F] than in the environment following [ʃ], since [ʃ] already has a lingual constriction. Speakers may be less likely to delete [u] following [F] (compared with [ʃ]) because deletion actually makes it harder to maintain devoicing.

6 Conclusion

Despite extensive past research on high vowel devoicing in Japanese, one issue that has remained open is whether the devoiced vowels are phonologically deleted or not. Following a previous study on this topic (Shaw & Kawahara, 2018b), the current EMA-based experiment explored this question with an extended stimulus set, with a new hypothesis that surrounding consonantal environments may modulate the deletion probability. The current experiment replicated the core findings of Shaw & Kawahara (2018b); there was a bimodal distribution in deletion probabilities for devoiced [u̥], with one mode representing vowels that fully retained their articulatory target and the other representing a linear tongue dorsum trajectory between flanking vowels.

The lack of a tongue dorsum height target, if it is due to vowel deletion, will presumably have phonological consequences for the language, including, at least, syllabification and syllable-based phonological patterns, e.g. accent placement and truncation patterns (as reviewed in the introduction). However, such evidence has not yet been identified. This could be for a number of reasons. The vowel may be retained, even if it lacks a tongue dorsum height target, affecting the phonetics in other dimensions. Possibilities include duration, lip movements, and the relative timing of flanking gestures. Alternatively, the vowel may be deleted while higher level prosodic structure, including moras and syllables, are retained, a possibility explored in Kawahara & Shaw (2018).

Additionally, the current experiment found an effect of fricative place of articulation on deletion probability—more deletion following [F] than [ʃ]—and individual differences in sensitivity to surrounding consonantal environments. These results are of descriptive importance, as we still know very little about the factors that condition variable phonological deletion of devoiced vowels in Japanese or, for that matter, any other languages that exhibit vowel devoicing. The current results highlight the importance of examining such behavior both within- and across- speakers, as sensitivity to phonological factors may also vary within a speech community.

Statement of Ethics

The current experiment was conducted with the approval of Western Sydney University and Keio University (Protocol number: HREC 9482). A consent form was obtained from each participant before the experiment.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions Statement

Designing the experiment: JS and SK; data analysis: JS; discussion of the results: JS and SK; writing up the paper: JS and SK.

Acknowledgments

Thanks to Chika Takahashi and Jeff Moore for help with the experiment, to Emily Grabowski for coding the data for devoicing, to Noah Macey for parsing out gestures. Thanks also go to Emily for work organizing the Matlab code for analysis into functions and a wrapper script, which are available on request. Portions of this study were presented at AMP 2018. This project is supported by JSPS grant #15F15715 to both authors. Thanks to two anonymous reviewers for comments.

References

Beckman, Mary. 1982. Segmental duration and the 'mora' in Japanese. *Phonetica* 39. 113–135.

Beckman, Mary. 1986. *Stress and non-stress accent*. Dordrecht: Foris.

Beckman, Mary & Atsuko Shoji. 1984. Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica* 41. 61–71.

Bennett, Ryan. 2020. Vowel deletion as phonologically-condition gestural overlap in Uspanteko. Talk presented at Keio-ICU LINC.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno & Paul Smolensky. 2008. Language universals in human brains. *Proceedings of the National Academu of Sciences* 5321–5325.

Berent, Iris, Donca Steriade, Tracy Lennertz & Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104. 591–630.

Brickhouse, C.J. & Kate Lindsey. 2020. Investigating the phonetics-phonology interface with field data: Assessing phonological specification through acoustic trajectories. Poster presented at the 96th meeting of the Linguistics Society of America.

Browman, Catherine & Louis Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3. 219–252.

Browman, Catherine & Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology* 6. 201–251.

Browman, Catherine & Louis Goldstein. 1992a. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Browman, Catherine & Louis Goldstein. 1992b. "targetless" schwa: An articulatory analysis. In G. Docherty & R. Ladd (eds.), *Papers in laboratory phonology II: Gesture, segment, prosody*, 26–56. Cambridge: Cambridge University Press.

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

Choi, John. 1995. An acoustic-phonetic underspecification account of marshallese vowel allophony. *Journal of Phonetics* 23. 323–347.

Cohn, Abigail. 1993. Nasalisation in English: Phonology or phonetics. *Phonology* 10. 43–81.

Cohn, Abigail. 2006. Is there gradient phonology? In Gisbert Fanselow, Caroline Fery, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 25–44. Oxford: Oxford University Press.

Cox, F & Janet Fletcher. 2017. *Australian english pronunciation and transcription*. Cambridge: Cambridge University Press.

Cutler, Anne, Takashi Otake & James McQueen. 2009. Vowel devoicing and the perception of spoke japanese words. *journal of the Acoustical Society of America* 1693.

Dauer, Rebecca M. 1980. The reduction of unstressed high vowels in Modern Greek. *Journal of the International Phonetic Association* 10(1-2). 17–27.

Delforge, AM. 2008. Gestural alignment constraints and unstressed vowel devoicing in andean spanish. *Proceedings of WCCFL* 26. 147–155.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier & Jacques Mehler. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25. 1568–1578.

Dupoux, Emmanuel, Erika Parlato, Sónia Frota, Yuki Hirose & Sharon Peperkamp. 2011. Where do illusory vowels come from? *Journal of Memory and Language* 64(3). 199–210.

Durvasula, K., H.-H. Huang, S. Uehara, Q. Luo & Y.-H. Lin. 2018. Phonology modulates the illusory vowels in perceptual illusions: evidence from Mandarin & English. *Laboratory Phonology* .

Faber, Alice & Timothy Vance. 2010. More acoustic traces of "deleted" vowels in Japanese. In Mineharu Nakayama & Carles Quinn (eds.), *Japanese/korean linguistics 9*, 100–113. CSLI.

Fais, Laurel, Sachiyo Kajikawa, Shigeaki Amano & Janet F. Werker. 2010. Now you hear it, now you don't: Vowel devoicing in japanese infant-directed speech. *Journal of Child Language* 37(2). 319–340.

Fujimoto, Masako. 2015. Vowel devoicing. In Haruo Kubozono (ed.), *The handbook of Japanese language and linguistics: Phonetics and phonology* 167-214, Mouton.

Fujimoto, Masako, Emi Murano, Seiji Niimi & Shigeru Kiritani. 2002. Difference in glottal opening pattern between tokyo and osaka dialect speakers: Factors contributing to vowel devoicing. *Folia Phoniatrica et Logopaedica* 54(3). 133–143.

Funatsu, Seiya & Masako Fujimoto. 2011. Physiological realization of japanese vowel devoicing. *Proceedings of Forum Acousticum* 2709–2714.

Garcia, D. 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics & Data ANalysis* 54(4). 1167–1178.

Gu, Chong. 2002. Smoothing spline anova models. R package.

Guenther, F. H., C. Y. Espy-Wilson, S. E. Boyce, M. L. Matthies, M. Zandipour & J. S. Perkell. 1999. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *journal of the Acoustical Society of America* 105. 2854–2865.

Hall, Kathleen Currie, Elizabeth Hume, Florian T. Jaeger & Andrew Wedel. 2018. The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4(S2).

Haraguchi, Shosuke. 1984. Some tonal and segmental effects of vowel height inJapanese. In M. Aronoff & R. T. Oehrle (eds.), *Language sound structure: Studies in phonology presented to morris halle by his teacher and students*, 145–156. Cambridge: MIT Press.

Hirayama, Manami. 2009. *Postlexical prosodic structure and vowel devoicing in Japanese*: University of Toronto Doctoral dissertation.

Imai, Terumi. 2004. *Vowel devoicing in Tokyo Japanese: A variationist approach*: Michigan State University Doctoral dissertation.

Imaizumi, Satoshi & A. Hayashi. 1995. Listener-adaptive adjustments in speech production: Evidence from vowel devoicing. *Annual Bulletin Research Institute of Logopedics and Phoniatrics,* 29. 43–48.

Imaizumi, Satoshi, A. Hayashi & T. Deguchi. 1995. Listener adaptive characteristics of vowel devoicing in Japanese dialogue. *journal of the Acoustical Society of America* 98. 768–778.

Ishihara, Shinichiro. 2011. Japanese focus prosody revisited: Freeing focus from prosodic phrasing. *Lingua* 121(13). 1870–1889.

Iskarous, Khalil, Joyce McDonough & Douglas H Whalen. 2012. A gestural account of the velar fricative in navajo. *Laboratory Phonology* 3(1). 195–210.

Isomura, Kazuhiro. 2009. *Onsei-wo oshieru [Teaching Japanese phonetics]*. Tokyo: Hitsuji Shobo.

Ito, Junko & Armin Mester. 1995. Japanese phonology. In John Goldsmith (ed.), *The handbook of phonological theory*, 817–838. Oxford: Blackwell.

Jaeger, Florian T. & Esteban Buz. 2018. Signal reduction and linguistic encoding. In Eva M. Fernández & Cairns Helen Smith (eds.), *The handbook of psycholinguistics*, 38–81. John Wiley & Sons.

Jain, Anil K. 1989. *Fundamentals of digital image processing*. Englewood Cliffs: Prentice Hall.

Jannedy, Stephanie. 1995. Gestural phasing as an explanation for vowel devoicing in Turkish. *Ohio State University Working Papers in Linguistics* 45. 56–84.

Johnson, Keith, Peter Ladefoged & Mona Lindau. 1993. Individual differences in vowel production. *journal of the Acoustical Society of America* 94(2). 701–714.

Jun, Sun-Ah & Mary Beckman. 1993. A gestural-overlap analysis of vowel devoicing in Japanese and Korean. Paper presented at the 67th annual meeting of the Linguistic Society of America, Los Angeles.

Jun, Sun-Ah, Mary Beckman & H.-J. Lee. 1998. Fiberscopic evidence for the influence on vowel devoicing of the glottal configurations for Korean obstruents. *UCLA Working Papers in Phonetics* 96. 43–68.

Kaneko, Ikuyo & Shigeto Kawahara. 2002. Positional faithfulness theory and the emergence of the unmarked: The case of kagoshima Japanese. *ICU English Studies 11* 5. 18–36.

Kawahara, Shigeto. 2015. A catalogue of phonological opacity in Japanese. *REPORTS of the Keio Institute of Cultural and Linguistic Studies* 46. 145–174.

Kawahara, Shigeto & Jason Shaw. 2018. Persistency of prosody. *Hana-bana: A Festshrift for Junko Ito and Armin Mester.* .

Kawahara, Shigeto, Jason A. Shaw & Shinichiro Ishihara. to appear. Assessing the prosodic licensing of wh-in-situ in japanese: A computational-experimental approach.

Natural Language and Linguistic Theory .

Kawakami, Shin. 1977. *Nihongo onsei gaisetsu [an overview of Japanese phonetics]*. Tokyo: Ohuusha.

Keating, Patricia A. 1988. Underspecification in phonetics. *Phonology* 5. 275–292.

Kibe, Nobuko. 2001. Sound changes in Kagoshima dialect. *Journal of the Phonetic Society of Japan* 5. 42–48.

Kilbourn-Ceron, Oriana & Morgan Sonderegger. 2018. Boundary phenomena and variability in Japanese high vowel devoicing. *Natural Language and Linguistic Theory* 36(1). 175–217.

Kilpatrick, Alexander, Shigeto Kawahara, Rikke Bungaard-Nielsen, Brett Baker & Janet Fletcher. 2020. Japanese perceptual epenthesis is modulated by transitional probability. *Language and Speech* .

Kondo, Mariko. 1997. *Mechanisms of vowel devoicing in Japanese*: University of Edinburgh Doctoral dissertation.

Kondo, Mariko. 2001. Vowel devoicing and syllable structure in Japanaese. In *Japanese/korean linguistics*, CSLI.

Kuriyagawa, F & Masayuki Sawashima. 1989. Word accent, devoicing and duration of vowels in Japanese. *Annual Bulletin of the Research Institute of Language Processing* 23. 85–108.

Lindblom, Björn. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35. 1773–1781.

Lindblom, Björn. 1990. Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech production and speech modeling*, 403–439. Dordrecht: Kluwer.

Maekawa, Kikuo. 1990. Production and perception of the accent in the consecutively devoiced syllables in Tokyo Japanese. *Proceedings of ICSLP 1990* .

Maekawa, Kikuo & H. Kikuchi. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: An interim report. In J. van de Weijer, Kensuke Nanjo & Tetsuo Nishihara (eds.), *Voicing in Japanese*, 205–228. Berlin: de Gruyter.

Martin, Andrew, Akira Utsugi & Reiko Mazuka. 2014. The multidimensional nature of hyperspeech: Evidence from Japanese vowel devoicing. *Cognition* 132(2). 216–228.

Matsui, Michinao. 2017. On the input information of the C/D model for vowel devoicing in Japanese. *Journal of the Phonetic Society of Japan* 21(1). 127–140.

McCarthy, John J. 2008. The gradual path to cluster simplification. *Phonology* 25(2). 271–319.

Moon, S. J. & Björn Lindblom. 1994. Interaction between duration, context and speaking style in English stressed vowels. *Journal of Acoustical Society of America* 96.1. 40–55.

Mücke, Doris, Martine Grice & Taehong Cho. 2014. More than a magic moment—Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics* 44. 1–7.

Munson, Benjamin, Jan Edwards, S.K. Shellinger, Mary E. Beckman & M.K. Meyer. 2010. Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical linguistics & phonetics* 24(4-5). 245–260.

Murray, Robert & Theo Vennemann. 1983. Sound change and syllable structure: Problems in Germanic phonology. *Language* 59. 514–28.

Murray, Robert W. 1988. *Phonological strength and early Germanic syllable structure*. München: Wilhelm Fink Verlag.

Myers, Scott. 1998. Surface underspecification of tone in chichewa. *Phonology* 15. 367–391.

Nakamura, Mitsuhiro. 2003. The articulation of vowel devoicing: A preliminary analysis. *On-in Kenkyuu [Phonological Studies]* 6. 49–58.

Nam, H., V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman & L. Goldstein. 2012. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America* 132(6). 3980–3989.

Nam, Hosung, Louis Goldstein, Elliot Saltzman & Dani Byrd. 2004. TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America* 115(5). 2430–2430.

Nielsen, Kuniko. 2015. Continuous versus categorical aspects of Japanese consecutive devoicing. *Journal of Phonetics* 52. 70–88.

Nogita, A., N. Yamane & S. Bird. 2013. The Japanese unrounded back vowel [W] is in fact rounded central/front [1/Y]. Paper presented at the Ultrafest VI, Edinburgh.

Ogasawara, Naomi. 2013. Lexical representation of Japanese vowel devoicing. *Language and Speech* 56(1). 5–22.

Perkell, J. S., M. L. Matthies, M. A. Svirsky & M. I. Jordan. 1993. Trading relations between tongue body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study. *JASA* 93. 2948–2961.

Pierrehumbert, Janet B. 1980. *The phonetics and phonology of English intonation*: MIT Doctoral dissertation.

Pierrehumbert, Janet B. & Mary Beckman. 1988. *Japanese tone structure*. Cambridge: MIT Press.

Poser, William. 1990. Evidence for foot structure in Japanese. *Language* 66. 78–105.

Rhodes, Richard. 1972. Cheyenne vowel devoicing and transderivational constraints. *Work Papers of the SUmmer Institute of Linguistics, University of North Dakota Session* 16. 52–55.

Roettger, Timo B. 2019. Researcher degree of freedom in phonetic research. *Laboratory Phonology* 10. http://doi.org/10.5334/labphon.147.

Saltzman, E.L. & Kevin. G. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecological psychology* 1(4). 333–382.

Sawashima, Masayuki. 1971. Devoicing of vowels. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics* 5. 7–13.

Shaw, Jason & Shigeto Kawahara. 2018a. Assessing surface phonological specification through simulation and classification of phonetic trajectories. *Phonology* 35. 481–522.

Shaw, Jason & Shigeto Kawahara. 2018b. The lingual gesture of devoiced [u] in Japanese. *Journal of Phonetics* 66. 100–119.

Shaw, Jason & Shigeto Kawahara. 2019. Effects of surprisal and entropy on vowel duration in Japanese. *Language and Speech* 62(1). 80–114.

Sjoberg, A.F. 1963. *Uzbek structural grammar*. Indiana University.

Smith, Caroline L. 2003. Vowel devoicing in contemporary French. *Journal of French Language Studies* 13(2). 177–194.

Starr, Rebecca L & Stephanie S. Shih. 2017. The syllable as a prosodic unit in Japanese lexical strata: Evidence from text-setting. *Glossa* .

Sugito, Miyoko & Hajime Hirose. 1988. Production and perception of accented devoiced vowels in Japanese. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics* 22. 19–36.

Tanner, James, Morgan Sonderegger & Francisco Torreira. 2019. Durational evidence that Tokyo Japanese vowel devoicing is not gradient reduction. *Frontiers in Psychology* .

Tiede, Mark. 2005. Mview. Software.

Tsuchida, Ayako. 1997. *Phonetics and phonology of Japanese vowel devoicing*: Cornell University Doctoral dissertation.

Vance, Timothy. 1987. *An introduction to Japanese phonology*. New York: SUNY Press.

Vance, Timothy. 2008. *The sounds of Japanese*. Cambridge: Cambridge University Press.

Vatikiotis-Bateson, Eric, Adriano Vilela Barbosa & Catherine T. Best. 2014. Articulatory coordination of two vocal tracts. *Journal of Phonetics* 44. 167–181.

Vennemann, Theo. 1988. *Preference laws for syllable structure and the explanation of sound change: With special reference to German, Germanic, Italian, and Latin*. Berlin: Mouton de Gruyter.

Vogel, Rachel. 2021. A unified account of two vowel devoicing phenomena: the case of Cheyenne. *Proceedings of Annual Meeting of Phonology* .

Watson, Catherine I. & Jonathan Harrington. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *journal of the Acoustical Society of America* 106(458-468).

Whang, James. 2018. Recoverability-driven coarticulation: Acoustic evidence from Japanese high vowel devoicing. *Journal of the Acoustical Society of America* 143. 1159–1172.

Whang, James. 2019. Effects of phonotactic predictability on sensitivity to phonetic detail. *Laboratory Phonology* 10(1).

Whang, James, Jason Shaw & Shigeto Kawahara. 2020. Acoustic consequences of vowel deletion in devoicing environments. Talk presented at LabPhon 17.

Yoshioka, H. 1981. Laryngeal adjustments in the production of the fricative consonants and devoiced vowels in Japanese. *Phonetica* 38. 236–251.

Zhang, M., C. Geissler & Jason Shaw. 2019. Gestural representations of tone in Mandarin: Evidence from timing alternations. *Proceedings of the 19th International Congress of Phonetic Sciences* 1803–1807.