The lingual articulation of devoiced /u/ in Tokyo Japanese

Jason Shaw* (Yale University)

Shigeto Kawahara (Keio University)

*Corresponding author: jason.shaw@yale.edu

New Haven, CT 06520, USA

Abstract

In Tokyo Japanese, /u/ is typically devoiced between two voiceless consonants. Whether the lingual vowel gesture is influenced by devoicing or present at all in devoiced vowels remains an open debate, largely because relevant articulatory data has not been available. We report ElectroMagnetic Articulography (EMA) data that addresses this question. We analyzed both the trajectory of the tongue dorsum across VC_1uC_2V sequences as well as the timing of C_1 and C_2 These analyses provide converging evidence that /u/ in devoicing contexts is optionally targetless—the lingual gesture is either categorically present or absent but seldom reduced. When present, the magnitude of the lingual gesture in devoiced /u/ is comparable to voiced vowel counterparts. Although all speakers produced words with and without a vowel height target for /u/, the frequency of targetlessness varied across speakers and items. The timing between C_1 and C_2, the consonants flanking /u/ was also effected by devoicing but to varying degrees across items. The items with the greatest effect of devoicing on this inter-consonantal interval were also the items with the highest frequency of vowel height targetlessness for devoiced /u/.

Keyword: Japanese, devoicing, /u/, articulatory phonetics, EMA, phonetic interpolation, Discrete Cosine Transform (DCT), Bayesian classifier, gestural coordination

General background

This paper examines the lingual articulation of devoiced /u/ in Tokyo Japanese. A classic description of the devoicing phenomenon is that high vowels are devoiced between two voiceless consonants and after a voiceless consonant before a pause (Fujimoto, 2015; Kondo, 1997, 2005; Tsuchida, 1997 among many others). This sort of description, high vowel devoicing in a particular context, applies to vowels in numerous other languages including e.g., French (Cedergren & Simoneau, 1985; Smith, 2003), Greek (Dauer, 1980; Eftychiou, 2010) Korean (Jun, Beckman, & Lee, 1998) and Uzbek (Sjoberg, 1963), but Tokyo Japanese is arguably the best studied case of vowel devoicing.

There is a large body of work on this phenomenon in Japanese, covering its phonological conditions (e.g., Kondo, 2005; Tsuchida, 1997), its interaction with other phonological phenomena like pitch accent (e.g., Kuriyagawa & Sawashima, 1989; Maekawa, 1990; Maekawa & Kikuchi, 2005; Vance, 1987) and prosodic structure (Kilbourn-Ceron & Sonderegger, 2017), its acoustic and perceptual characteristics (Beckman & Shoji, 1984; Faber & Vance, 2000; Matsui, 2014; Nielsen, 2015; Sugito & Hirose, 1988), and studies of the vocal folds (Fujimoto, Murano, Niimi, & Kiritani, 2002; Hirose, 1971; Sawashima, 1971; Tsuchida, 1997). Fujimoto (2015) provides a recent, comprehensive overview of this research. While now we have a good understanding of many aspects of high vowel devoicing in Tokyo Japanese, there is little data available on the lingual gestures of high vowels when they are devoiced. The only study that we are aware of is Funatsu & Fujimoto (2011), which used EMMA (ElectroMagnetic

Midsagittal Articulography) with concurrent imaging of the vocals fold using nasal endoscopy. They found little difference between devoiced and voiced /i/ in terms of lingual articulation. However, this experiment used only one speaker and one item pair (/kide/ vs. /kite/). The study included four repetitions of each item, and offered no quantitative analyses of the data. Our study is intended to expand on this previous work by reporting more data from more speakers and more extensive quantitative analysis.

Why is it important to study the lingual gestures of devoiced vowels? There are a few lines of motivation behind the current study. First, consider Figure 1, taken from Fujimoto et al's (2002) study, which used nasal endoscopy to image the glottal gestures of high vowel devoicing in Japanese.

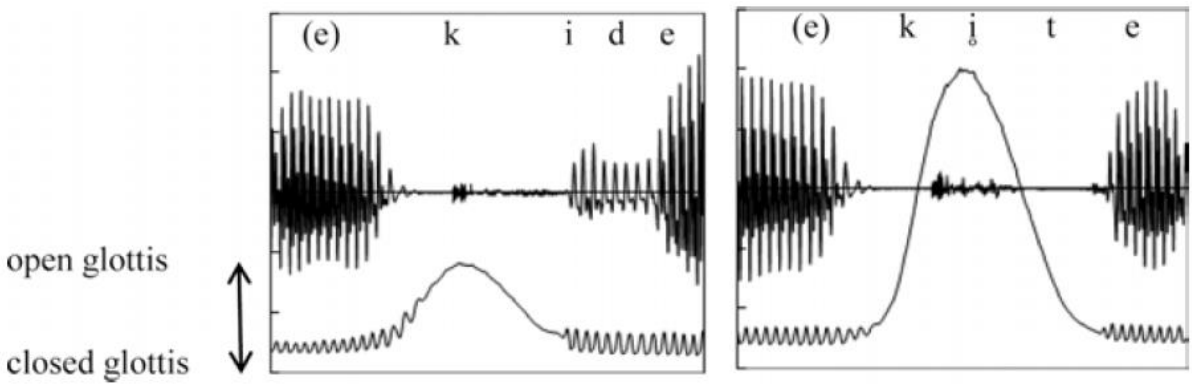


Figure 1: The degrees of glottal abduction in Japanese. The left panel: a voiceless stop followed by a voiced stop, which has a single abduction gesture for /k/. The right panel: a voiceless stop /k/ followed by a voiceless vowel and another voiceless stop /t/, which also has a single abduction gesture. The magnitude of the abduction gesture in the right panel is larger than twice the size of the abduction gesture in the left panel. Taken from Fujimoto et al. (2002), cited and discussed in Fujimoto (2015).

Figure 1 shows that a Japanese devoiced vowel has a single laryngeal gesture of greater magnitude than a single consonant gesture, or even the sum of two voiceless consonant gestures (c.f., Munhall & Lofqvist, 1992 for English which shows the latter pattern).[1] This observation implies that Japanese devoiced vowels involve active laryngeal abduction, not simply overlap of two surrounding gestures. This conclusion in turn implies that Japanese speakers exert *active* laryngeal control over devoiced high vowels (cf. Jun & Beckman 1993 for an analysis that relies on passive gestural overlap, to be discussed below). To the extent that Japanese speakers actively control the laryngeal gesture for devoiced vowels, are lingual gestures of devoiced vowels also actively controlled? There are competing views on this matter. On the one hand, it seems logical that active control of a non-contrastive property (allophonic devoicing) would imply active control of a contrastive property (tongue position in the vocal tract). On the other hand, the way that devoicing operates physiologically in Japanese obliterates much of the acoustic signature of lingual articulation. Speakers may not exert active control over aspects of articulation that do not have salient auditory consequences.

The second line of motivation for the current study is the question of whether "devoiced" vowels are simply devoiced or deleted. This issue has been discussed extensively in previous studies of Japanese high vowel devoicing. Kawakami (1977: 24-26) argues that vowels delete in some environment and devoice in others, but he offers no phonological or phonetic evidence. Vance (1987) raised and rejected the hypothesis that high vowels in devoicing contexts are

[1] Munhall and Lofqvist (1992) investigate the timing and magnitude of laryngeal gestures in the consonants /s/ and /t/ in the sequences *kiss ted* spoken at different speech rates. At slow speech rates two distinct laryngeal gestures can be identified but at faster speech rates the gestures merge into one laryngeal gesture approximating the sum of the two smaller consonantal gestures.

deleted. Kondo (2001) argues that high vowel devoicing is actually deletion based on a phonological consideration. Devoicing in consecutive syllables is often prohibited (although there is much variability: Nielsen 2015), and Kondo argues that this prohibition stems from a prohibition against complex onset or complex coda (i.e. *CCC). On the other hand, Tsuchida (1997) and Kawahara (2015) argue that bimoraic foot-based truncation (Poser, 1990) counts a voiceless vowel as one mora (e.g. [sɯ̥to] from [sɯ̥toraiki] 'strike', *[stora]).[2] If /u/ was completely deleted losing its mora, the bimoraic truncation should result in *[stora]. Hirayama (2009) makes a similar phonological argument by showing that devoiced vowels' moras are just as relevant for Japanese *haiku* poetry as moras in voiced vowels. However, just because moras for the devoiced vowels remain does not necessarily mean that the vowel is present. The adjacent consonant could conceivably host the mora and syllable—this hypothesis is actually proposed by Matsui (2014), who argues that Japanese has consonantal syllables in this environment (see Dell and Elmedlaoui, 2002 for similar analyses of Tashlhiyt Berber and Moroccan Arabic). Thus, evidence for either deletion or devoicing from a phonological perspective is mixed (see Fujimoto 2015: 197-198 for other studies addressing this debate).[3]

Previous acoustic studies show that on spectrograms, vowels leave no trace of lingual articulation except for coarticulation on surrounding consonants, which lead them to conclude that vowels are deleted (Beckman 1982; Beckman & Shoji 1984; Whang 2014). An anonymous

[2] Here and throughout we use the symbol [u] to refer to a broad phonetic transcription. The actual realizations of this vowel in our data in Tokyo Japanese more generally tend not to be as back or as rounded as [u] is strictly defined in the IPA. See Vance (2008:51) for a detailed description of Japanese /u/. We return to this point when discussing our specific hypotheses below.

[3] Tsuchida (1997) argues that there is "phonological devoicing" as well as "phonetic devoicing".

reviewer questions this finding reported in past work. In principle, a change in the sound source, from modal voicing to turbulence, is independent of the resonance properties of the vocal tract (e.g., Stevens, 1999). We might therefore expect to be able to identify formant structure in the aperiodic energy characteristic of devoiced vowels. The reported absence of such structure in past studies may follow from the particular location of turbulent energy sources excited in the vocal tract preceding devoiced vowels in Japanese and their perseverative influence on devoiced vowels. Most of the voiceless consonants preceding devoiced vowels in Japanese are fricatives or affricates that involve turbulence generated by a narrow channel of air in the anterior portion of the vocal tract.[4] The formants produced by these consonants are resonances of the cavity in front of the noise source, i.e., the front cavity. Using ElectroPalatoGraphy (EPG), Matsui (2014) shows that the narrow channel characteristic of fricatives and affricates persists across following devoiced vowels. The perseveration of the airflow channel across vowels no doubt helps to sustain devoicing by maintaining high intraoral air pressure, but may also contribute to the obliteration of spectral cues to vowel articulation. In contrast, when the energy source is modal phonation at the glottis, higher formants (above F1) of vowels result from (coupled) resonances of both the front and back cavities, which provide very different acoustic signatures from resonance of the front cavity alone. These factors may contribute to the claim that devoiced vowels show no acoustic traces of lingual articulation. Beckman (1982: 118, footnote 3) states that "deletion" is a better term physically, because "there is generally no spectral evidence for a voiceless vowel", whereas "devoicing" is a better term psychologically,

[4] Japanese lacks singleton /p/, except in some recent loanwords, and /t/ is affricated before high vowels, so the only stop consonant conditioning devoicing that does not involve turbulence generated in the anterior portion of the vocal tract is /k/.

because Japanese speakers hear a voiceless vowel even in the absence of spectral evidence (c.f., Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999). Beckman and Shoji (1984: 64) likewise state that "[w]hen the waveform of a devoiced syllable is examined, however, neither its spectral nor its temporal structure indicates the presence of a voiceless vowel." These statements embrace the "deletion" view, at least at the speech production level. Even if vowel devoicing involves phonological deletion, or deletion of some component (feature, gesture) of the vowel, it could be the case that its application is optional or variable, influenced by various linguistic and sociological factors (Fujimoto, 2015; Imaizumi & Hayashi, 1995; Nielsen, 2015).

Not all phonetic studies have embraced the deletion view, however. The clearest instantiation of an alternative view is the "gestural overlap theory" of high vowel devoicing (Faber & Vance, 2000; Jun & Beckman, 1993; Jun et al., 1998). In this theory, high vowel devoicing occurs when glottal abduction gestures of the surrounding consonants overlap with the vowel's glottal gesture (though cf. Figure 1). In this sense, the high vowel devoicing processes in Japanese (and other languages like Korean) are "not…phonological rules, but the result of extreme overlap and hiding of the vowel's glottal gesture by the consonant's gesture" (Jun & Beckman 1993: p.4). This theory implies that there is actually no deletion—oral gestures remain the same, but do not leave their acoustic traces because of devoicing.

To summarize, there is an active debate about whether the lingual gestures of "devoiced" vowels in Japanese are present but inaudible due to devoicing or absent altogether, possibly because of phonological deletion. Vance (2008), the most recent and comprehensive phonetic textbook on Japanese, states that this issue is not yet settled. Studying lingual movements of devoiced vowels will provide crucial new evidence. Recall that the only past study on this topic, Funatsu & Fujimoto (2011), is based on a small number of tokens and one speaker. In this

paper, we expand the empirical base, reporting more repetitions (10-15) of ten real words produced by six naive speakers, and we deploy rigorous quantitative methods of analysis, as detailed by Shaw and Kawahara (submitted-a). While Shaw and Kawahara (submitted-a) focused on motivating the computational methodology, this paper reports more data and examines the consequences of vowel devoicing for the temporal organization of gestures. In particular, we examine C-C timing across devoiced vowels in order to ascertain whether devoicing impacts the gestural coordination of flanking consonants.

Hypotheses

Building on the previous studies reviewed in this section, we entertain four specific hypotheses about the lingual articulation of devoiced vowels, stated in (1)

(1) Hypotheses about the status of lingual articulation in devoiced vowels

H1: **full lingual targets**—the lingual articulation of devoiced vowels is the same as for voiced counterparts.

H2: **reduced lingual targets**—the lingual articulation of devoiced vowels is phonetically reduced relative to voiced counterparts.

H3: **targetless**—devoiced vowels have no lingual articulatory target.[5]

H4: **optional target**—devoiced vowels are sometimes targetless (deletion is optional, token-by-token).

[5] We use the term "targetless" rather than "deletion", because the latter term commits to (1) a surface representation, (2) a process mapping an underlying representation to a surface representation and (3) the identity of the underlying representation. Our experiment is solely about the surface representation, and hence the term "targetless" is better.

The passive devoicing hypothesis, or the gestural overlap theory (e.g. Jun & Beckman 1993), maintains that there is actually no phonological deletion, and hence would predict that lingual gestures would remain intact (=H1). This is much like Brownman and Goldstein's (1992) argument that apparently deleted [t] in *perfect memory* in English keeps its tongue tip gesture. This is also the conclusion that Funatsu & Fujimoto (2011) reached for devoiced /i/ in their sample of EMMA data. Besides the small sample size mentioned above, another caveat is concurrently collected nasal endoscopy may have promoted stable lingual gestures across contexts, since retraction of the tongue body may trigger a gag-reflex.

Even if devoiced high vowels are not phonologically deleted, it would not be too surprising if the lingual gestures of high vowels were phonetically reduced, hence H2 in (1). At least in English, more predictable segments tend to be phonetically reduced (Aylett & Turk, 2004, 2006; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Jurafsky, Bell, Gregory, & Raymond, 2001). In Japanese, the segmental identity of devoiced vowels is often highly predictable from context (Beckman & Shoji, 1984; Whang, 2014). Due to devoicing, moreover, the acoustic consequences of a reduced lingual gesture would not be particularly audible to listeners. Hence, from the standpoint of effort-distinctiveness tradeoff (Hall, Hume, Jaeger, & Wedel, 2016; Lindblom, 1990), it would not be surprising to observe reduction of oral gestures in high devoiced vowels.

The phonological deletion hypothesis, which was proposed by various authors reviewed above, predicts that there should be no lingual targets for devoiced vowels (=H3 in (1)). However, we know that many if not all phonological patterns are variable, i.e., optional (e.g., Coetzee & Pater, 2011), to some degree. There is an increasing body of evidence that phonological and

phonetic patterns are stochastic (Hayes & Londe, 2006; McPherson & Hayes, 2016; Pierrehumbert, 2001). For example, Bayles, Kaplan and Kaplan (2016) have shown recently that there is intra-speaker variation in French schwa production such that the same speaker may produce a word with or without a schwa vowel. Even studies using high spatio-temporal resolution articulatory data capable of picking up gradience have revealed that some phonological patterns, e.g., place assimilation, can be optional within a speaker (Ellis & Hardcastle, 2002; Kochetov & Pouplier, 2008). Therefore, we need to consider the possibility that deletion of lingual gestures in devoiced vowels is optional, by assessing the presence/absence of phonetic specification on a token-by-token basis (=H4).

In contrast to H1 and H2, H3/H4 present some thorny methodological challenges. Even when we limit ourselves to one phonetic dimension, distinguishing categorical absence of phonetic specification from heavy phonetic reduction is challenging. To pursue this challenge, we adopt an assumption in the literature on phonetic underspecification (e.g., Keating, 1988) that the absence of phonetic specification in some dimension results in phonetic interpolation between flanking segments. Making use of this assumption, the essence of our approach is to assess the linearity of the trajectory between flanking vowels, a method introduced in Shaw & Kawahara (submitted-a). To implement this analysis with appropriate baselines on a token-to-token basis, we setup Bayesian classifiers of our devoiced vowel tokens based on two categories of training data: voiced vowels in similar contexts (to our devoiced test items) and linear trajectories between flanking segments. The classifier returns the probability that each token belongs to the voiced vowel trajectory as opposed to the linear interpolation trajectory. This provides us with a rigorous quantitative approach to assessing both degree differences in phonetic reduction and the absence of phonetic specification in a particular phonetic dimension.

Our analysis focuses on the phonetic dimension that is most characteristic of the target vowel. For the case of /u/, the focus of this study, that dimension is tongue height. Although we follow Vance (2008) and many other authors in the Japanese phonetics literature in using the symbol /u/, the backness and rounding components of this vowel in Japanese are not precisely as the IPA symbol implies. In Japanese, /u/ is rather central, as reported in the ultrasound study of Nogita et al. (2013) and shown in the instructional MRI images of Isomura (2009). Vance (2008: 55) calls Japanese /u/ the hardest of the Japanese vowels to describe, in part because of the labial component, which involves compression instead of rounding. In contrast to /u/ in other languages, e.g., English, where the longitudinal position of EMA sensors placed on the lips gives a reasonable measure of rounding (Blackwood-Ximenes, Shaw, & Carignan, 2017), lip compression associated with /u/ in Japanese is difficult to detect from sensors on the vermillion borders of the lips. Hence, when it comes to differentiating trajectories of voiced vowels from devoiced vowels, the labial and backness components of Japanese /u/ cannot be expected to provide a particularly strong signal against the backdrop of natural variability in vowel production. For this reason, we focus on tongue height and designed stimulus materials in which the target vowel /u/ is always flanked by non-high vowels, e.g., /…eCuCo…/. If the tongue body does not rise from its position for the mid-vowel /e/ to /u/ before falling to /o/, but proceeds instead on a linear trajectory from /e/ to /u/, we would conclude that the token lacks a vowel height target for /u/.

With respect to the targetless hypotheses (H3/H4), we acknowledge that lacking a phonetic height target for /u/ is not quite the same as vowel deletion, as intended by some of the studies reviewed above. Certainly, vowel deletion implies targetlessness in the height dimension but it also implies that other phonetic dimensions are similarly targetless. It is conceivable that a vowel that lacks a height target for /u/ is nonetheless specified in other phonetic dimensions,

e.g., a laryngeal gesture, tongue backness, lip compression, etc., in which case targetless is an inappropriate designation for the vowel. On the other hand, our materials are well-designed to falsify H3/H4 if it can be shown that the movement trajectory of the tongue does not approximate linear interpolation (see Browman & Goldstein, 1992a for a comparable approach). Our study is also capable of falsifying H1 by showing systematic differences between tongue trajectories in voiced and voiceless tokens and H2 by showing either no difference between lingual movements or a bimodal distribution representing presence and absence of a height target. In this way, the limitations in focusing on the vowel height dimension extend to our assessment of all four hypothesis. Our analysis of C1-C2 timing provides additional evidence bearing on H1-H4.

Besides adjudicating between the hypotheses in (1), another line of motivation for this study stems from the insight that devoiced vowels may provide on how laryngeal and supra-laryngeal gestures are coordinated. As our review of the literature revealed, there is evidence that some aspects of devoiced vowels—the laryngeal gesture and the airflow channel—may be phonetically controlled to ensure devoicing. This suggests that passive devoicing may not be the right synchronic analysis of the Japanese facts, but it is unclear what implications controlled devoicing has for the organization of lingual gestures. The laryngeal gestures that contribute to devoicing ostensibly originate not with the vowel but with the flanking consonants. Nevertheless, EGG evidence in Figure 1 (Fujimoto et al., 2002) indicates that the timing and magnitude of laryngeal gestures shift when vowel devoicing is at stake. When there is an intervening high vowel, the laryngeal gestures of flanking voiceless consonants aggregate to form one large-magnitude laryngeal abduction centered on the vowel. Aggregation of C_1 and C_2 laryngeal gestures across the vowel may also require the oral gestures associated with the consonants to come closer together, if the internal temporal integrity of the consonants, i.e. the

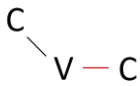
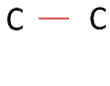
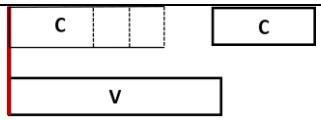
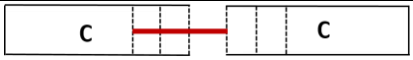
articulatory binding of oral and laryngeal gestures (Kingston, 1990), is to be maintained. However, it may be that articulatory binding is violable. In this case, Japanese may sacrifice the internal timing of consonants in order to maintain temporal coordination between supra-laryngeal gestures. On this scenario, supra-laryngeal gestures maintain patterns of temporal organization generally characteristic of VC_1uC_2V sequences, even as laryngeal gestures aggregate to devoice /u/. To assess these possibilities, we also analyzed the timing of the flanking consonants, C_1 and C_2, and the inter-consonantal interval, i.e., the interval from the release of C_1 to the achievement of target of C_2 (specific measurements are defined below).

The conflict between maintaining consonant-internal timing between laryngeal and oral gestures, on the one hand, and between maintaining inter-gestural timing across consonants and vowels, on the other hand, is complicated by the possibility that the devoiced vowel is absent (H3) at least in some tokens (H4). We assume, following, e.g., Smith (1995), that in a VC_1VC_2V sequence in Japanese, C_2 is coordinated locally to the preceding vowel. If that vowel is absent, yielding C_1C_2V, we expect C-C coordination, whereby C_2 is coordinated with C_1.

One way to assess the presence of a coordination relation is to evaluate the predicted covariation between temporal intervals (Shaw et al., 2011). Our Japanese materials afford this opportunity. To make the comparison between C-V and C-C timing more concrete, we express the alternatives as coordination topologies (in the sense of Gafos, 2002; Gafos & Goldstein, 2012) in Table 1 along with consequences for relative timing. The rectangles represent activation durations for gestures. The dotted lines indicate variation in intra-gestural activation duration. Under C-V coordination, we assume here that the start of consonant and vowel gestures are coordinated in time, i.e., the gestures are in-phase (Goldstein, Nam, Saltzman, & Chitoran, 2009). Under this coordination regime, shortening of C_1 would expose more of the inter-

consonantal interval (ICI), predicting a negative correlation between C_1 duration and ICI. Under C-C coordination, on the other hand, the end of C_1 is coordinated with the start of C_2. Variation in activation duration for C_1 impacts directly when in time C_2 begins. Hence, under C-C coordination no trade-off between C_1 duration and ICI is predicted. By assessing the predicted covariation between C_1 and the inter-consonantal interval, we can adjudicate between C-V and C-C timing, potentially providing an independent argument for the presence/absence of the vowel (H3/H4). Evidence for C-V timing is also evidence that the vowel is present to some degree, whereas C-C timing is indirect evidence that the vowel is absent, or at least that it is not specified to the extent that it affects the coordination of flanking consonants. This is particularly useful since our analysis of vowel target presence/absence is limited to the height dimension. The additional analysis, although indirect, provides another way to assess whether the vowel is phonetically specified.

Table 1: Schematic illustration of C1 duration variation under different coordination topologies.

	C-V coordination	C-C coordination
Coordination Topology		
Temporal intervals		

Through the set of analyses described above, we aim to evaluate the lingual articulation of the devoiced vowel and the consequences of devoicing for the temporal organization of flanking segments, i.e., the consonants that define the devoicing environment.

Methodology

Speakers

Six native speakers of Tokyo Japanese (3 male) participated. Participants were aged between 19 and 22 years at the time of the study. 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 naïve to the purpose of the experiment. They were compensated for their time and local travel expenses. In subsequent discussion we refer to the speakers as S01 through S06, numbered in the order in which they were recorded.

Materials

A total of 10 target words, listed in Table 2, were included in the experiment. These included five words containing /u/ in a devoicing context (second column) and a set of five corresponding words with /u/ in a voiced context (third column). The target /u/ is underlined in each word. Together, these 10 words constitute minimal pairs or near minimal pairs. The word pairs are matched on the consonant that precede /u/. They differ in the voicing specification of the consonant following /u/. The consonant following /u/ is voiceless in devoicing context words (second column) and voiced in voiced context words (third column). This study focused on /u/ and did not consider /i/ for several practical reasons, the most important one being that collecting both /u/ and /i/ tokens would mean reducing the repetitions for each target word and the analytical approach we planned (following Shaw and Kawahara, submitted-a) requires a

large number of repetitions per word. We focused on /u/ rather than /i/, because the former is more likely to be devoiced, as confirmed by the study of high vowel devoicing using the Corpus of Spontaneous Japanese (Maekawa & Kikuchi 2005; see also Fujimoto et al. 2015 and references cited therein).

Table 2: stimulus items. W and K show the environments in which Kawakami (1971) and Whang (2014) predict deletion.

Comments	Devoicing/deletion	Voiced vowel
V deletion (K, W)	ɸusoku 不足 'shortage'	ɸuzoku 付属 'attachment'
V devoicing (K, W)	ʃutaise: 主体性 'subjectivity'	ʃudaika 主題歌 'theme song'
V deletion (K, W)	katsutoki 勝つ時 'when winning'	katsudo: 活動 'activity'
V devoicing (K) V deletion (W)	hakusai 白菜 'white cabbage'	jakuzai 薬剤 'medicine'
V deletion (K, W)	masutaa マスター 'master'	masuda 益田 'Masuda (a surname)'

The consonant preceding /u/ draws from the set: /s/, /k/, /ʃ/, /ts/, /ɸ/.[6] All of these consonants may contribute to the high vowel devoicing environment, when they occur as C_1 in $C_1V_{[High]}C_2$ sequences, but some of them are also claimed to condition deletion—not just devoicing—of the following vowel in the same environment. According to Kawakami (1977), /s/, /ts/, and /ɸ/ condition deletion of the following /u/, while /k/ and /ʃ/ condition devoicing only, although recall that he offers no phonological or phonetic evidence. According to Whang (2014), vowel deletion occurs when the identity of the devoiced vowel is predictable from context. His recoverability-based theory predicts that /u/ will be deleted following /s/, /ts/, /ɸ/, and /k/ but

[6] Although here and throughout we list /ts/ and /ɸ/ in slashes, we note that they are largely (but not entirely) predictable allophones: [ts] is the allophone of /t/ that occurs before /u/; [ɸ] is the allophone of /h/ that occurs before /u/.

that /u/ will be present (though devoiced) following /ʃ/. The predictions match Kawakami's intuition for four of the five consonants in our stimuli (/s/, /ts/, /ɸ/, /ʃ/). The point of divergence is the /k/ environment. Whang's theory predicts vowel deletion following /k/, whereas Kawakami claims that the vowel is present (although devoiced) following /k/. The predictions for the stimulus set from Whang (2014) are labelled as "(W)" in the first column of Table 2; those due to Kawakami are labelled "(K)"; converging predictions are labelled as "(W,K)".

We did not include stimuli in which high vowels are surrounded by two sibilants, as it is known that devoicing may be inhibited in this environment (Hirayama 2009; Fujimoto 2015; Maekawa & Kikuchi 2005; Tsuchida 1997). In addition, if the vowel is followed by /h/, /ɸ/, or /ç/ devoicing may be inhibited (Fujimoto 2015). Our stimuli avoided this environment as well.

We avoided any words in which the vowel following the target vowel is also high, because consecutive devoicing is variable (Fujimoto, 2015; Nielsen, 2015). We also chose near minimal pairs in such a way that accent always matches within a pair. More specifically, /u/s in /hakusai/ and /jakuzai/ are both accented, meaning that the pitch fall begins at the /u/, but all the other target /u/s are unaccented, and they carry low pitch. Although intonational accents can influence vowel coarticulation, at least in English (Cho, 2004), Tsuchida (1997) shows that young Japanese speakers, at the time of 1997, show no effects of pitch accent on devoicing, so that

controlling for accent is important but may not be crucial.[7] All the stimulus words are common words.[8]

The target words were displayed in the carrier phrase: *okee* _____ *to itte* 'Okay say _____'. The preceding word *okee* was chosen, as it ends with a vowel /e/, which differs in height from the target vowel, /u/. Participants were instructed to speak as if they were making a request of a friend.

Each participant produced 10-15 repetitions of the target words, generating a corpus of 690 tokens for analysis. We aimed to get 15 tokens from each speaker, but if a sensor came off in the late stages of the experiment, after we had collected at least 10 repetitions of all target words, we ended the session. Words were presented in Japanese script (composed of hiragana,

[7] There are very little if any durational differences between accented and unaccented vowels (Beckman, 1986), which would otherwise potentially affect the deletability of /u/. Since accented /u/ is not longer than unaccetend /u/, this is yet another reason not to be too concerned about the placement of accent.

[8] The written frequencies of our stimulus items in the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a 104.3 million word corpus of diverse written materials (Maekawa et al., 2014), are as follows: /ɸu̲soku/ (6,342), /ɸu̲zoku/ (2,273), /ʃu̲taise:/ (4,480), /ʃu̲daika/ (1,584), /katsu/ (6,163), /katsudo:/ (31,440), /masutaa/ (1,517), /masuda/ (0), /hakusai/ (639), and /jakuzai/ (1,561). The only item with low frequency in the corpus is /masuda/, probably because it is a proper noun, although this word is not uncommon as a Japanese name.

katakana and kanji characters as required for natural presentation) and fully randomized with 10 additional filler items that did not contain /u/.

Equipment

The current experiment used an NDI Wave electromagnetic articulograph system sampling at 100 Hz to capture articulatory movement. The NDI wave tracks fleshpoints with an accuracy typically within 0.5 mm (Berry, 2011). NDI wave 5DoF sensors were attached to three locations on the sagittal midline of the tongue, and on the upper and lower lips near the vermillion border, lower jaw (below the 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. The most posterior sensor, henceforth TD, was attached as far back as was comfortable for the participant, ~4.5-6 cm. A third sensor, henceforth TB, was placed on the tongue body roughly equidistant between the TT and TD sensors. Figure 2 illustrates the location of the lingual sensors for one participant. Acoustic data were recorded simultaneously at 22 kHz with a Schoeps MK 41S supercardioid microphone (with Schoeps CMC 6 Ug power module).

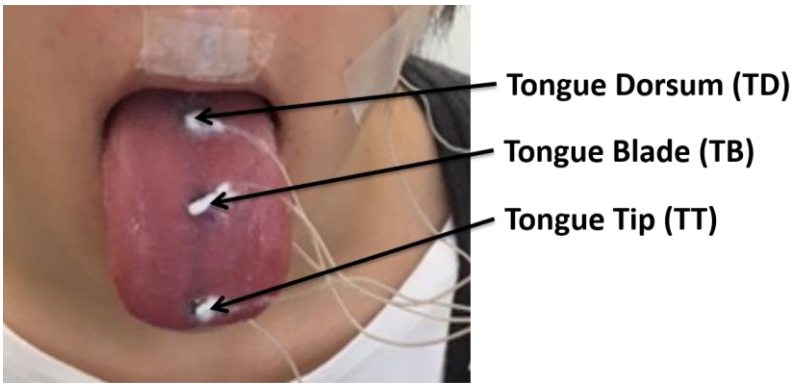


Figure 2: a representative illustration of lingual sensor placement for one participant.

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 allowed for online monitoring of hesitations, mispronunciations and disfluencies. These were rare, but when they occurred, the experimenter repeated the trial. Participants were instructed to read the target word in the carrier phrase fluently, as if providing instructions to friend, and told explicitly not to pause before the target word. Each trial consisted of a short (500 ms) preview presentation of the target word followed by presentation of the target word within the carrier phrase. The purpose of the preview presentation was to further facilitate fluent reading of the target word within the carrier phrase, since it is known that a brief visual presentation of a word facilitates planning (Davis et al., 2015) and, in particular, to discourage insertion of phonological phrase boundary between “okee (okay)” in the carrier phrase and the target word.

Post-processing

Following the main recording session, we also recorded the occlusal 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, left/right mastoid and nasion sensors, and the three sensors on the occlusal plane. The head-corrected data was rotated so that the origin of the spatial coordinates corresponds to the occlusal plane at the front teeth. All articulatory signals were smoothed using Garcia’s robust smoothing algorithm (Garcia, 2010).

Analysis

Presence of voicing

One of the authors and a research assistant each went through the spectrograms and waveforms of all the tokens, and confirmed that /u/ in the devoicing environments are all devoiced (Figure 5 below provides a sample spectrogram), whereas /u/ in the voicing environments was voiced. This is an unsurprising result, given that vowel devoicing is reported to be obligatory in the normal speech style of Tokyo Japanese speakers (Fujimoto 2015).[9]

Lingual targets

All stimulus items were selected so that the vowels preceding and following the target /u/ were non-high. In order to progress from a non-high vowel to /u/, the tongue body must rise. For some stimulus items, e.g., /ɸu̲soku/, /ɸu̲zoku/, /ʃu̲taise:/, /ʃu̲daika/, the tongue body may also retract from the front position required for /e/ in the carrier phrase (*okee* ____ *to itte*) to the more posterior position required for /u/. The degree to which the tongue body retracts for /u/, i.e., the degree to which /u/ is a back vowel, has been called into question, with some data suggesting that /u/ in Japanese is central (Nogita, Yamane, & Bird, 2013), similar to "fronted" variants of /u/ in some dialects of English (e.g., Blackwood-Ximenes et al., 2017; Harrington, Kleber, & Reubold, 2008). We therefore focus on the height dimension, in which /u/ is

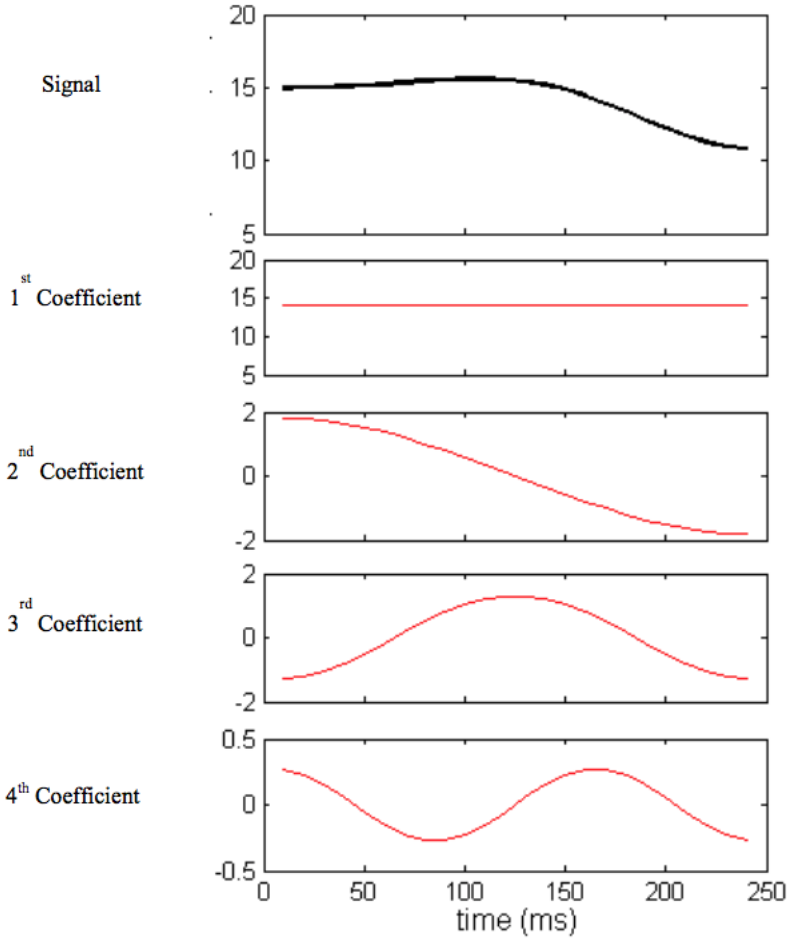
[9] Though see Maekawa & Kikuchi (2005) who show that devoicing may not be entirely obligatory in spontaneous speech—in their study, overall, /u/ is devoiced about 84% of the time in the devoicing environment. However, as Hirayama (2009) points out, their study is likely to contain environments where there are two consecutive high vowels, which sometimes resist devoicing (Kondo 2001; Nielsen 2015).

uncontroversially distinct from /o/. As an index of tongue body height, we used the TD sensor, the most posterior sensor of the three sensors on the tongue, which provides comparable data to past work using fleshpoint tracking to examine vowel articulation (Browman & Goldstein, 1992a; Johnson, Ladefoged, & Lindau, 1993).

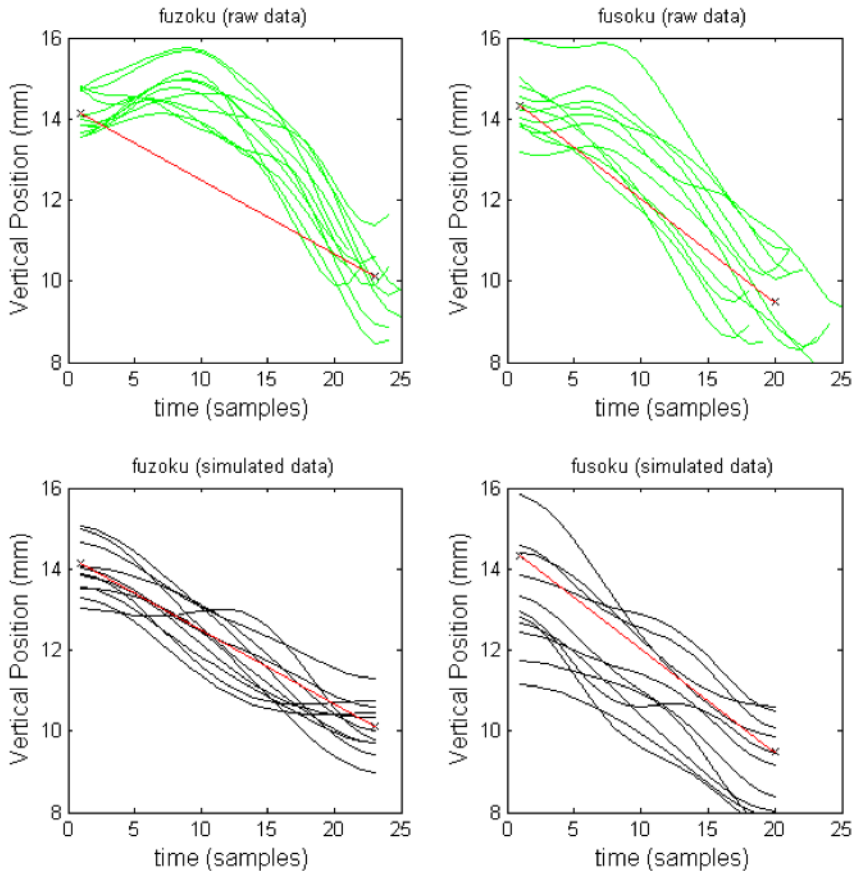
Our analytical framework makes use of the computational toolkit for assessing phonological specification proposed by Shaw and Kawahara (submitted-a). This framework evaluates presence/absence of an articulatory target based upon analysis of continuous movement of the tongue body across $V_1C_1\underline{u}C_2V_3$ sequences (Figure 3). Analysis involves four steps: (1) fit Discrete Cosine Transform (DCT) components to the trajectories of interest; (2) define the targetless hypothesis based on linear movement trajectories between the vowels flanking /u/; (3) simulate "noisy" linear trajectories using variability observed in the data, in which the means are taken from the DCT coefficients of the targetless trajectory, and the standard deviations are taken from the observed data; (4) classify voiceless tokens as either "vowel present" or "vowel absent" based on comparison to the "vowel present" training data (voiced vowels) and the "vowel absent" training data (based on linear interpolation), using a Bayesian classifier.

Figure 3: Steps illustrating the computation analysis (based on Shaw & Kawahara, submitted-a)

(A) Step 1: Fitting four DCT components to a trajectory. The top panel is the raw signal of /V1C1uC2V3/. The rest of the panels shows each DCT components.



(B) Steps 2 & 3: Define a linear interpolation (shown as red lines) and generate a noisy null trajectory (the bottom two panels), based on the variability found in the raw data (the top two panels). Left=trajectories for /eɸuzo(ku)/; right=trajectories for /eɸuso(ku)/.



(C) Step 4: Train a Bayesian classifier in terms of DCT coefficients. "Vowel present" defined in terms of voiced vowels. "Vowel absent" defined in terms of the linear trajectory.

$$p(T|Co_1, Co_2, Co_3, Co_4) = \frac{p(T)\, p(Co_1, Co_2, Co_3, Co_4|T)}{p(Co_1, Co_2, Co_3, Co_4)}$$

where

T = *targetless (linear interpolation), target present (voiced vowel)*
Co_1 = *1st DCT Coefficient*
Co_2 = *2nd DCT Coefficient*
Co_3 = *3rd DCT Coefficient*
Co_4 = *4th DCT Coefficient*

Shaw and Kawahara (submitted-a) demonstrate that four DCT coefficients are sufficient for representing TD height trajectories over VCuCV intervals with an extremely high degree of precision ($R^2 > .99$). They show moreover that each DCT coefficient has a plausible linguistic interpretaation: the 1st coefficient generally picks out general TD height across the analysis interval, the 2nd coefficient captures V1-V3 movement, the 3rd coefficient picks out movement associated with /u/, if any, and the 4th coefficient captures additional coarticulatory effects from surrounding consonants.

We view the potential for interpreting DCT components as signal modulations associated with linguistically relevant units, i.e., gestures, as an interesting difference from other times series analyses (e.g., GAMMs, functional data analysis, SSANOVA). Nevertheless, we would like to stress that, although it seems clear that the magnitude of the /u/ gesture is related to the 3rd DCT component, we cannot demonstrate that the 3rd DCT coefficient is picking out all and only the rising movement associated with /u/. For example, an increased magnitude of /u/ may also lead

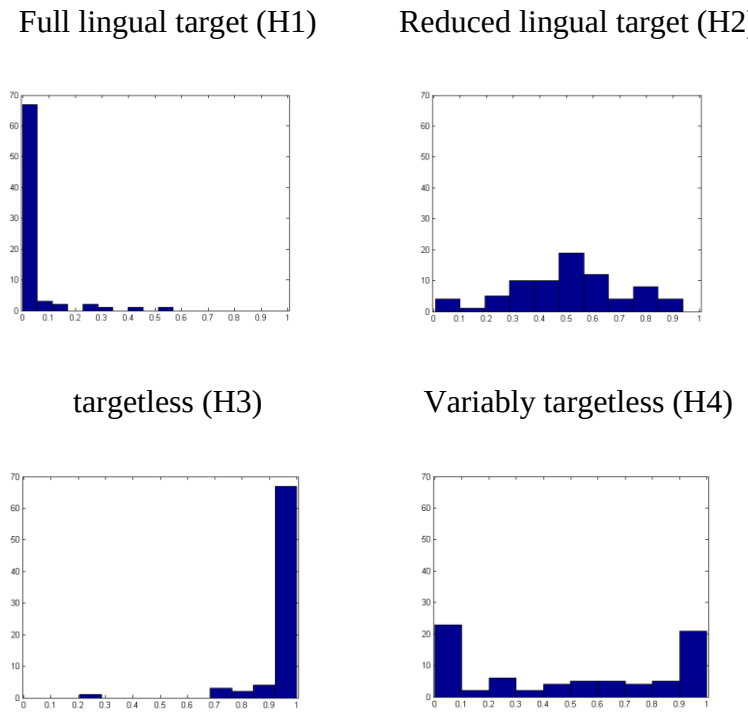
to an increase in the 1st DCT coefficient, which is related to the average trajectory height. For this reason, we took what we believe to be the most conservative approach and based our classification of trajectories on all four DCT coefficients. In this way, the properties of DCT that are most pertinent to our analysis are the compression property—four DCT coefficients provide a very close approximation to the raw trajectories—and the statistical independence of the coefficients, an assumption of the naïve Bayes classifier that is met by DCT coefficients. We report the classification results for each devoiced token in terms of the posterior probability of targetlessness, i.e. the likelihood that the trajectory follows a linear interpolation between V1 and V3 instead of rising like the tongue body does in voiced tokens of /u/.

We acknowledge that our choice to model the entire VC_1uC_2V trajectory has the consequence that any differences in TD height associated with the voicing specification of C_2 will also factor into the results. Consonants that contrast minimally in voicing are known to have some differences in lingual articulation, owing in part to aerodynamic factors associated with voicing. In Japanese, voiceless stops and affricates tend to have tighter constrictions than voiced stops, as indicated by greater lingual-palatal contact (Kochetov & Kang, 2017) while the pattern is reversed for fricatives—voiced fricatives tend to have more contact than voiceless fricatives, at least in the anterior portion of the palate (Nakamura, 2003). In our items, C_2 was always a coronal consonant. In this precise environment (C_2 following a devoiced consonant), Nakamura (2003) reports EPG data showing a difference in palate contact between Japanese /z/ and /s/ (more contact for /z/) but only in the anterior portion of the palate. There were no differences at the tongue dorsum, the focus of our analysis.

The four hypotheses introduced in (1) can each be evaluated by examining the distribution of posterior probabilities across tokens. Figure 4 presents hypothetical distributions corresponding

to each of the hypotheses. If devoiced vowels have full lingual targets (H1), then we expect the probability of targetless to be low, as is shown in the top left panel of Figure 4. This figure was made by submitting items with voiced /u/ to the classifier. If H2 is correct, and devoiced vowels have reduced lingual targets, then the distribution of posterior probabilities should be centered around .5, as in the top right panel. This figure was made by submitting DCT components half the magnitude of voiced vowels to the classifier. If devoiced vowels lack lingual targets altogether (H3), then the distribution of posterior probabilities should be near 1.0, as in the bottom left panel of the figure. This figure was made by submitting DCT coefficients from noisy linear trajectories (simulated) to the classifier. Lastly, if lingual targets are variably present, as in H4, then we expect to see a bimodal distribution, with one mode near 0 and the other near 1.0, as shown in the bottom right panel. This figure was made by submitting a mix of tokens from sampled from linear trajectories and from voiced vowels to the classifier.

Figure 4: Four hypothetical posterior probability patterns. The vertical axis of each histogram shows posterior probabilities generated by the Bayesian classifier summarized in Figure 3(C). The histogram in the top left panel was obtained by submitting the /ɸuzoku/ (voiced vowel) tokens to the Bayesian classifier. The histogram in the bottom left was obtained by submitting the same number of simulated linear interpolation "vowel absent" trajectories to the classifier. The top right panel was generated by stochastic sampling of DCT coefficients that were averaged between "target present" (H1) and "target absent" (H3) values. The right bottom panel was created by sampling over both targetless and full vowel target tokens.



An anonymous reviewer points out that the velar stop preceding /u/ in /hakusai/ complicates the interpretation of our classification results for this item. In our other target items, the

devoiced /u/ is immediately preceded by coronal, /ʃ/, /ts/, /s/, or labial, /ɸ/, consonants, which do not dictate a large rise in tongue dorsum (TD) height. For these items, we interpret a rise in TD height as progress towards the goal of /u/ production. More precisely, our analytical approach assesses whether the height of the devoiced /u/ trajectory is closer to a linear trajectory between flanking vowels, i.e., no rise at all, or the TD rise observed in a voiced vowel in the same consonantal environment. In /hakusai/, the devoiced /u/ is preceeded by a velar stop, /k/. The TD rises from the first /a/ towards the target of /k/ before falling again for the following /a/ (see Figure 6, fourth row from the top). The large TD rise for the /k/ immediately preceding the target /u/ may obscure differences in vowel height specification across voiced and devoiced /u/. Specifically, the TD trajectory in the /akusa/ portion of /hakusai/ may be more similar to the TD trajectory of the /akuza/ portion of /jakuzai/ than to a linear interpolation between /a/ and /a/ because of the influence that /k/ exerts over the TD in both /akusa/ and /akuza/. For this reason, we have excluded /hakusai/ from the classification analysis. The raw TD trajectories from /hakusai/ and /jakuzai/ are reported and these data are included in the analyses of inter-consonantal timing described below, as this analysis does not require that we assess the influence of /u/ on the TD signal.

Consonantal timing across devoiced vowels

In addition to examining the continuous trajectory of the tongue dorsum, we also investigated the timing of consonants preceding and following /u/. If the consonants flanking /u/ are coordinated in time with the lingual gesture for the vowel (e.g., see Smith, 1995 for a concrete proposal), reduction or deletion of that vowel gesture may perturb consonant timing. In this way, consonant timing offers another angle on how devoicing influences lingual articulation.

For this analysis, we identified articulatory landmarks from consonants flanking /u/ based on the primary oral articulator, e.g., tongue tip for /t/, /s/, tongue blade for /ʃ/, tongue dorsum for /k/, lips for /ɸ/, etc, for each gesture. The affricate /ts/ in /katsutoki/ and /katsudou/ was treated as a single segment. We determined the start and end of consonantal constrictions with reference to the velocity signal in the movements toward and away from constriction. Data were displayed in Mview and articulatory landmarks were extracted using *findgest*, an Mview labeling procedure (Tiede, 2005). Both the start of the constriction, a.k.a. the achievement of target of the consonant, and the end of the constriction, a.k.a. the release landmark, were extracted at the timepoint corresponding to 20% of peak velocity in the movement towards/away from consonantal constrictions, a heuristic applied extensively in other recent (Bombien, Mooshammer, & Hoole, 2013; Gafos, Hoole, Roon, & Zeroual, 2010; Marin, 2013; Marin & Pouplier, 2010; Shaw, Chen, Proctor, & Derrick, 2016; Shaw, Gafos, Hoole, & Zeroual, 2011; Shaw, Gafos, Hoole, & Zeroual, 2009). As a first pass, we used all three spatial dimensions of the positional signal and corresponding tangential velocities to parse consonantal landmarks. This approach was practicable for a majority of the consonant tokens in our corpus. However, for some tokens, parsing landmarks based on the tangential velocity was problematic. One issue was that the amount of spatial displacement associated with a consonant gesture was sometimes too small to produce a prominent velocity peak. This occurred most often for the tongue tip movement from /s/ to /d/ in /masuda/. When the velocity peak is very small it cannot reliably delineate controlled movement. Consonant tokens that could not be parsed from the velocity signal were excluded from analysis (25 tokens, 3.6% of the data). Another issue was that, in some tokens, there were not distinct tangential velocity peaks associated with

the release of C1 and the movement towards C2. This is because movement towards C2 in one dimension, such as advancement of the tongue for /t/ in /ʃutasise:/ or /d/ in /ʃudaika/ overlapped in time with movement in another dimension associated with C1, such as lowering of tongue for /ʃ/. For many of these cases, we were able to isolate distinct velocity peaks for C1 and C2 by focusing on the primary spatial dimension of movement for the gesture, such as lowering for the release of /ʃ/ and raising toward the target for /t/. This approach, suggested in *Guidelines for using Mview* (Gafos, Kirov, & Shaw, 2010) allowed us to consider a greater number of tokens for analysis. The parseability of consonants based on tangential velocity (as opposed to component velocity) was unrelated to whether C2 was voiced or voiceless. The total number of tokens parsed by tangential vs. component velocities is provided by item in the Appendix.

Figure 5 provides an example of consonantal landmarks parsed for for /ʃ/ and /t/, the consonants flanking target /u/ in /ʃutaise:/ based on movement in the vertical dimension only. The vertical black lines indicate the timestamp of the consonantal landmarks. They extend from the threshold of the velocity peak used as a heuristic for parsing the consonant up to the corresponding positional signal. The dotted line in the velocity panels extends from 0 (cm/s), or minimum velocity.

The interval between landmarks, labeled C_1 for /ʃ/ and C_2 for /t/, are the consonant constriction durations. The interval between the consonants, or inter-consonantal interval (ICI), defined as the achievement of target of C_2 minus the release of C_1 (see also, inter-plateau interval in Shaw

& Gafos, 2015) was also analyzed. At issue is whether this interval varies with properties of the lingual gesture for the intervening /u/.

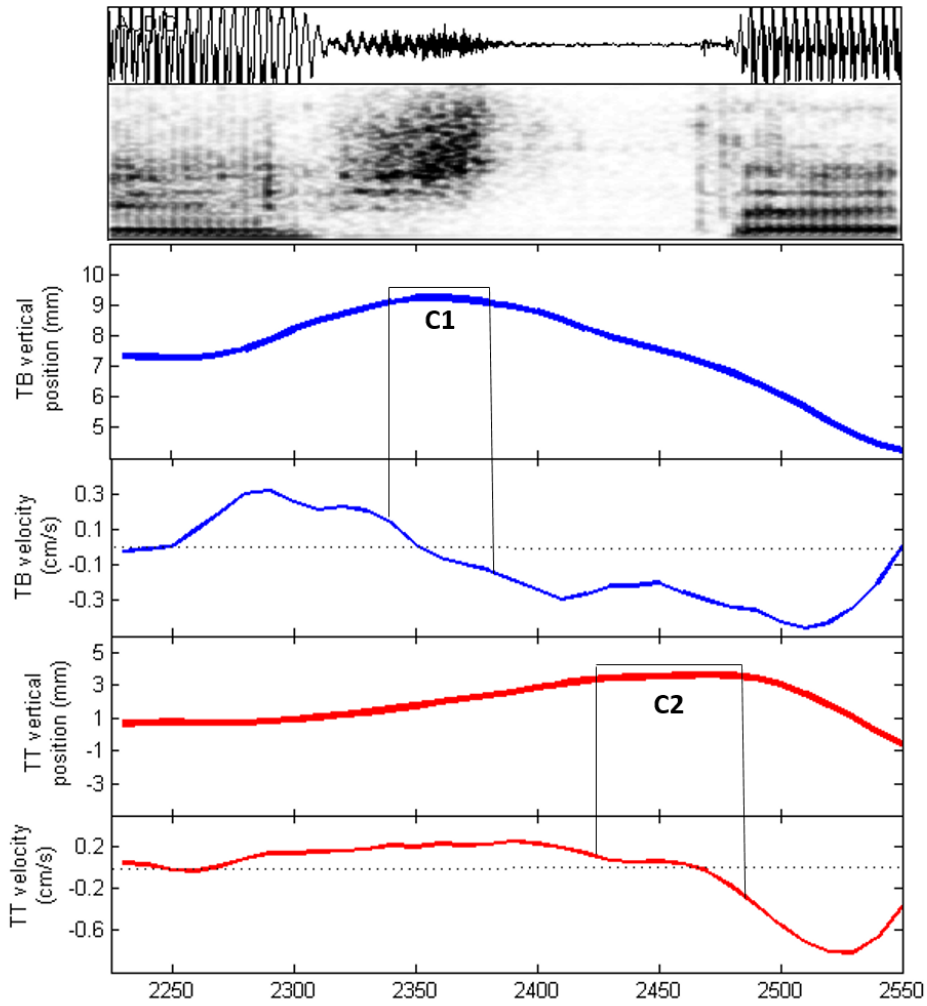


Figure 5: An illustration of how consonantal landmarks, /ʃ/ (C1) and /t/ (C2), were parsed from a token of /ʃutaise:/. The thick blue line shows the vertical position of the tongue blade (TB); the thin blue line shows the corresponding velocity signal. The thick red line shows the vertical position of the TT; the thin red line shows the corresponding velocity signal. Consonant onsets and offsets were based on a threshold of peak velocity, vertical black lines, in the movements toward and away from target.

Data exclusion

The tongue tip sensor became unresponsive for S04 on the sixth trial. We think that this was due to wire malfunction, possibly due to the participant biting on the wire. The tongue tip trajectory is relevant only for the analysis of consonant timing. Due to missing data, the consonant timing analysis for this speaker is based on only 5 trials. Analyses involving other trajectories are based on 11 trials for this participant.

Results

Presence/absence of articulatory height targets

Figure 6 summarizes the data on TD height across speakers and words. Each panel shows TD height (*y*-axis) over time (*x*-axis) for tokens of a target word with a voiced vowel (blue lines) and devoiced counterpart (red lines). The columns show data from different speakers and the rows show the different voiced-devoiced dyads. An interval of 350 ms (35 samples of data), beginning with the vowel preceding the target, is shown in each panel. Despite the variation in speech rate (note, for example, that the rise of the TD for /k/ at the right of panels displaying /ɸu̲soku/~/ɸu̲zoku/ in the top row is present to different degrees across speakers), a 350 ms window is sufficient to capture the three-vowel sequence including the target /u/ and preceding and following vowels for all tokens of all words across all speakers.

To facilitate interpretation of the tongue height trajectories, annotations are provided for the first speaker (leftmost column of panels). The trajectories begin with the vowel preceding the target vowel, e.g., /e/ from the carrier phrase in the case of /ɸu̲soku/ and /ʃu̲taise:/, /a/ in /ka̲tsutoki/, etc. Movements corresponding to the vowel following /u/ are easy to identify—since the vowel following /u/ is always non-high, it corresponds to a lowering of the tongue.

The label for the target vowel, /u/, has been placed in slashes between the flanking vowels. We note also that the vertical scales have been optimized to display the data on a panel-by-panel basis and are therefore not identical across all panels. In particular, across speakers the TD is lower in the production of *masuda~masutaa* than for many of the other dyads and that this influences the scale for most speakers (S02-S06).

Differences between voiced and devoiced dyads (red and blue lines, respectively, in the figure) include cases in which the tongue tends to be higher in the neighborhood of /u/ for the voiced than for the devoiced member of the dyad. The top left panel, /ɸu̲soku/ for S01, exemplifies this pattern (a zoom-in plot is provided in Figure 7). In this panel, the blue lines rise from /e/ to /u/ while the devoiced vowel trajectory is a roughly linear trajectory between /e/ and /o/. Our analysis assesses this possibility specifically by setting up stochastic generators of the competing hypothesis, lingual target present (based on the voiced vowel) vs. lingual target absent (based on linear interpolation), and evaluates the degree to which the voiceless vowel trajectory is consistent with these options.

Figure 6: vertical TD trajectories of all speakers, all items. The

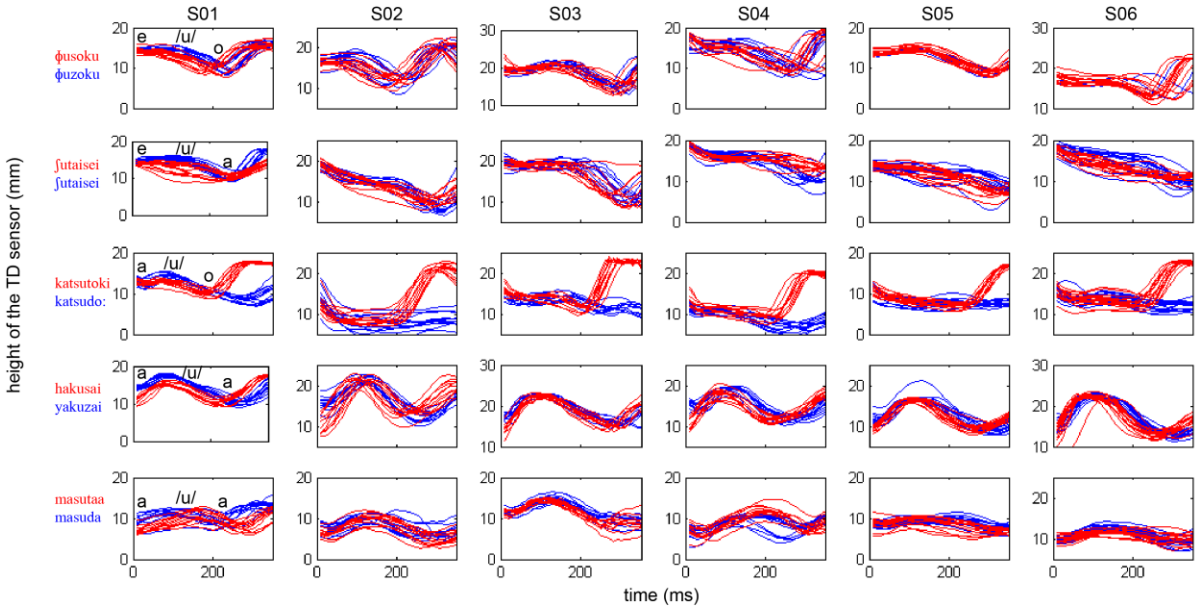


Figure 7: TD trajectories of /ɸusoku/ (red) and /ɸuzoku/ (blue) for S01. The blue lines generally show a rise from the /e/ in the carrier phrase to the /u/ before lowering again to the target for /o/. Some red lines show a similar rise from /e/ to /u/; others follow a downward cline from /e/ to /o/ without rising for /u/. The computational analysis described above classifies each red line as either belonging to the voiced category (blue lines) or to a linear trajectory from /e/ to /o/.

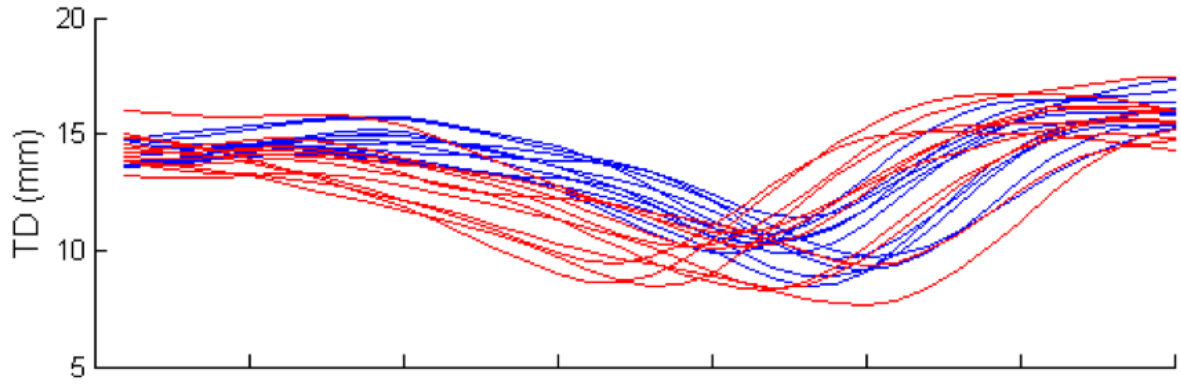


Table 3 provides the average posterior probabilities across tokens by speaker and by word. Since our analysis involves a stochastic component—the linear trajectory corresponding to the vowel absent scenario is stochastically sampled—we repeated the simulation and classification steps multiple times. The standard deviations of the posterior probabilities across 1000 repetitions are given in parenthesis. The probability of targetlessness varies across speakers rather widely, from speakers that have a high probability of targetless vowels, e.g. 0.70 for S01, to speakers with a much lower probability of targetlessness, e.g., 0.23 for S05. There are also differences across items. The targetlessness probability is highest for /ʃu̲taise:/ (0.69) and /katsutoki/ (0.60) with lower probabilities for /ɸu̲soku/ (0.43) and /masutaa/ (0.36).

Table 3: Posterior probability of lingual targetlessness (vowel height)

	S01	**S02**	**S03**	**S04**	**S05**	**S06**	**average**
ɸu̲soku	0.47(.11)	0.38(.14)	0.75(.07)	0.85(.08)	0.01(.01)	0.11(.06)	**0.43**
ʃu̲taise:	0.92(.03)	0.66(.17)	0.81(.06)	0.92(.07)	0.05(.05)	0.80(.15)	**0.69**
katsu̲toki	0.70(.22)	0.23(.11)	0.81(.13)	0.93(.03)	0.13(.11)	0.78(.13)	**0.60**
masu̲taa	0.73(.20)	0.11(.08)	0.04(.05)	0.02(.02)	0.74(.20)	0.52(.14)	**0.36**
average	**0.70**	**0.35**	**0.60**	**0.68**	**0.23**	**0.55**	

Figure 8 shows histograms of posterior probabilities both across speakers (left panels) and within speakers (right panels). Probabilities close to 1 indicate a linear trajectory, i.e., no rise in TD height for /u/, while probabilities near 0 indicate that the trajectory for the devoiced vowels resembles the trajectory for voiced vowels. As illustrated earlier (Figure 4), each of the hypotheses about lingual articulation of devoiced vowels motivated from the literature makes distinct predictions about the shape of these histograms. Figure 8 shows that, for all words, the distribution of probabilities is distinctly bimodal. This indicates that many of the tokens of

devoiced vowels in our corpus were either produced with a full lingual target or produced as linear interpolation between flanking vowels. Only a small number of tokens are intermediate, i.e., posterior probabilities in the range of .5 indicating reduction relative to voiced vowels but not to the degree that would result in linear interpolation. The bimodal distributions support the "optional targetless hypothesis" (=H4) (see also Figure 5), at least for the population of tokens drawing from the six speakers in this study.

The right panels of Figure 8 present posterior probabilities by speaker. These figures illustrate that both inter- and intra-speaker variation contribute to the bimodality of posterior probabilities. Consider /ʃu̲taise:/. Of the six speakers in the study, three of them, S01, S03, S04, show strong tendencies towards a linear trajectory, i.e., an /u/ lacking a height target. For these three speakers, the vast majority of tokens have targetless probabilities greater than .9. Two others, S02 and S06, show slightly weaker tendencies towards targetlessness. The majority of the tokens for these two speakers are between .7 and .9, values that indicate these tokens are still much closer to the linear trajectory than to the voiced vowel. Only one speaker, S05, shows a strong tendency towards producing a full vowel height target. This speaker's data contributes most heavily to the mode near 0 in the group data. Of the 15 tokens of /ʃu̲taise:/ with targetless probabilities of less than .1, 10 tokens are from S05. Thus, part of the bimodality in the group data derives from inter-speaker differences: speaker S05 tends to produce full vowel height targets in devoiced vowels; the other five speakers do not. However, we can see that optionality within speakers, i.e., intra-speaker variation, also contributes somewhat to bimodality in the group data. Speakers S01, S03, and S04, those who showed the strongest tendency towards targetlessness, indicated by the peaks in the histograms between .9 and 1, all produced one token that is at the other end of the histogram, indicating that it resembles the height trajectory of the voiced vowel. The other items tell similar stories vis the contribution of both inter- and

intra-speaker variation to bimodality. Overall, it seems that bimodality in the group data derives both from across speaker variation in the probability with which /u/ is produced with a height target as well as intra-speaker variation. All six speakers demonstrated that the linear trajectory for /u/ is within their production repertoire, producing at least one item with high probability of targetlessness.

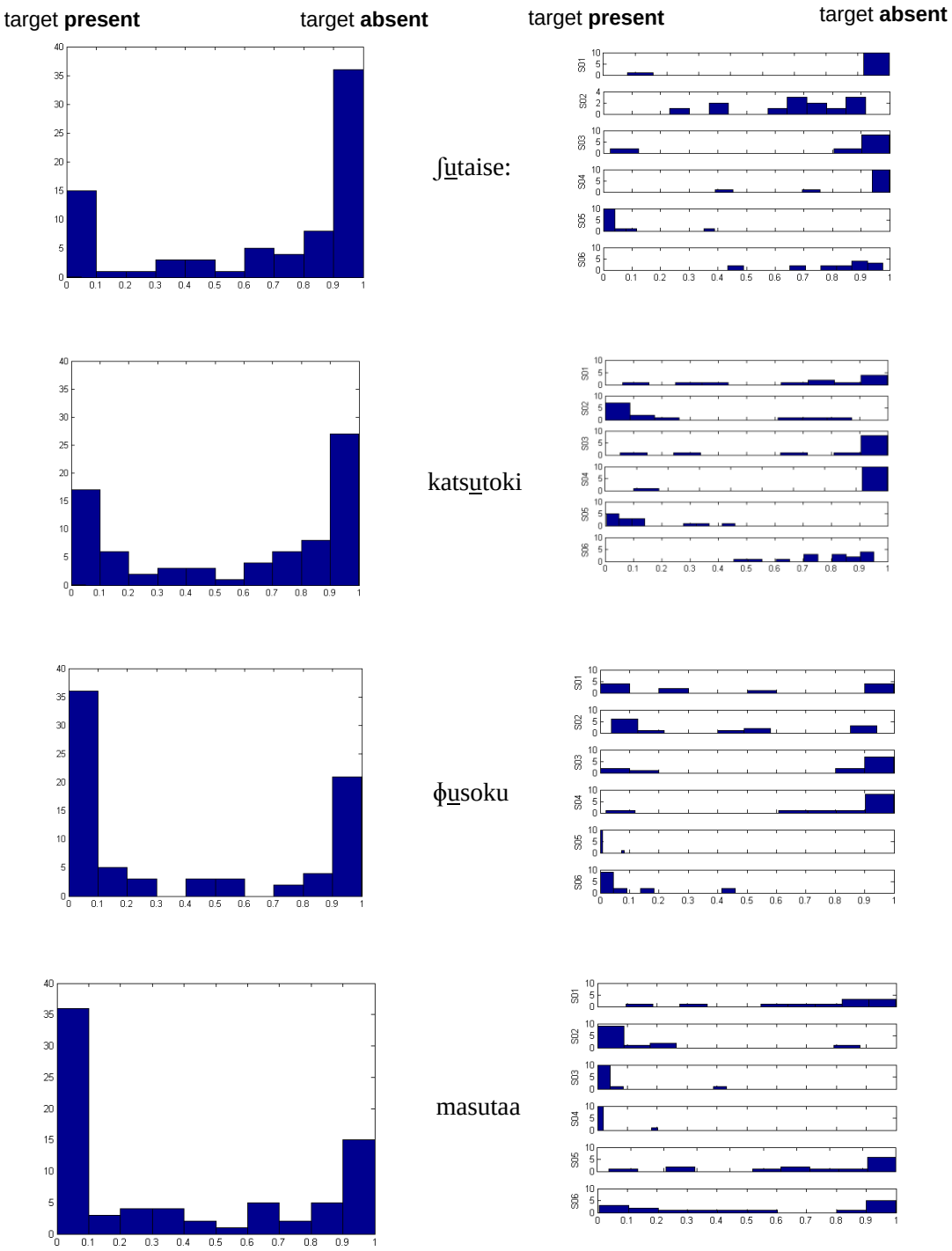


Figure 8: Posterior probability of targetlessness for each token organized by item. The left panels aggregate across speakers; the right panels show probabilities for each speaker from S01 (top) to S06 (bottom). All items show a roughly bimodel pattern anchored by tokens with high probabilities of targetlessness (numbers close to 1) on the right side of the figures and tokens with low probabilities of targetlessness (numbers close to 0) on the left side of the figures.

The inter-consonantal interval

We next turn to the timing of the consonants flanking /u/, asking whether devoicing influences relative timing between the preceding and following consonants. Past work on laryngeal control of devoiced vowels indicates that the laryngeal gestures associated with flanking consonants aggregate to form one large laryngeal gesture near the center of the vowel (Fujimoto et al. 2002; Fujimoto 2015; see Figure 1). Here, we investigate the consequences of this laryngeal reorganization for the oral gestures associated with the consonants. A decrease in the inter-consonantal interval, defined as the interval spanning from the release of C_1 to the achievement of target of C_2, across devoiced vowels relative to voiced vowels would indicate that laryngeal reorganization "pulls" the oral gestures of the consonants closer together in time. Alternatively, a consistent ICI across voiced and devoiced vowels would indicate that articulatory binding (consonant-internal temporal organization) of oral and laryngeal gestures is perturbed to achieve devoicing.

Alongside our interest in the ICI interval as an indication of how oral gestures respond to laryngeal reorganization, given the results above on vowel height targets, we can also ask whether ICI is impacted by the presence/absence of a lingual height target for the intervening vowel. We opted to analyze the data in terms of this categorical difference (instead of using the raw probabilities as predictors) because, as shown in the histograms in Figure 8, the data are largely categorical in nature. We applied the Bayesian decision rule, interpreting (targetless) probabilities greater than .5 as indicating that the vowel height target was absent and probabilities less than .5 indicate that the target was present.

Figure 9 summarizes the inter-consonantal interval (ICI) across words containing a voiced vowel and words containing a devoiced vowel. Since ICI is a difference, it is possible for it to

be negative. This can happen when C2 achieves its target before the release of C1, as commonly observed in English consonant clusters (e.g., Byrd, 1996). The average ICI is around zero for *ɸusoku~ɸuzoku* indicating that the lips remain approximated until the tongue blade achieves its target, a temporal configuration which does not at all interfere with the achievement of a height target for the vowel. The /ɸs/ sequence of consonants is a front-to-back sequence, in that the place of articulation for the labial fricative, the first consonant, is anterior to the place of articulation of the alveolar fricative, the second consonant. Consonant clusters with a front-to-back order of place tend to have greater overlap than back-to-front clusters (Chitoran, Goldstein, & Byrd, 2002; Gafos, Hoole, et al., 2010; Wright, 1996; Yip, 2013). For other dyads, the consonantal context requires longer average ICI's, with median values ranging from ~50 ms to ~170 ms. The variation reflects the broader fact about Japanese that consonantal context has a substantial influence on the duration of the following vowel (see Shaw & Kawahara, submitted-b for a large scale corpus study).

With respect to how devoicing influences ICI, Figure 9 shows that the effect of vowel devoicing on ICI varies across consonantal environments. For *ʃutaise:~ʃudaika* and *katsutoki~katsudo"*, and to a lessor degree *hakusai~jakuzai*, ICI is longer when the vowel is voiced. For *ɸusoku~ɸuzoku*, ICI is quite similar across words and for *masutaa~masuda*, the presence of vowel voicing actually results in shorter ICI.

To assess the statistical significance of the trends shown in Figure 9, we fitted a series of three nested linear mixed effects models to the ICI data using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in R. The baseline model included word DYAD as a fixed factor, since it is clear that ICI depends in part on the identity of the particular consonants, and random intercepts for speaker, to account for speaker-specific influences, e.g., speech rate, on ICI. The

second model added VOWEL VOICING to the baseline model as a fixed factor. The third model included the interaction between DYAD and VOWEL VOICING.

Table 4 summarizes the model comparison. Adding VOWEL VOICING to the model leads to significant improvement over the baseline, indicating a general trend for shorter ICI across devoiced vowels. However, as we observed, VOWEL VOICING affects ICI for some dyads positively and for other dyads negatively. Adding the interaction between DYAD and VOWEL VOICING leads to further improvement over the model with DYAD and VOWEL VOICING as non-interacting fixed factors. These results indicate that the trends observed in Figure 9 are statistically reliable. ICI varies depending on whether the vowel is voiced or devoiced; how it varies depends also on the word or possibly on the identity of the flanking consonants within the word.

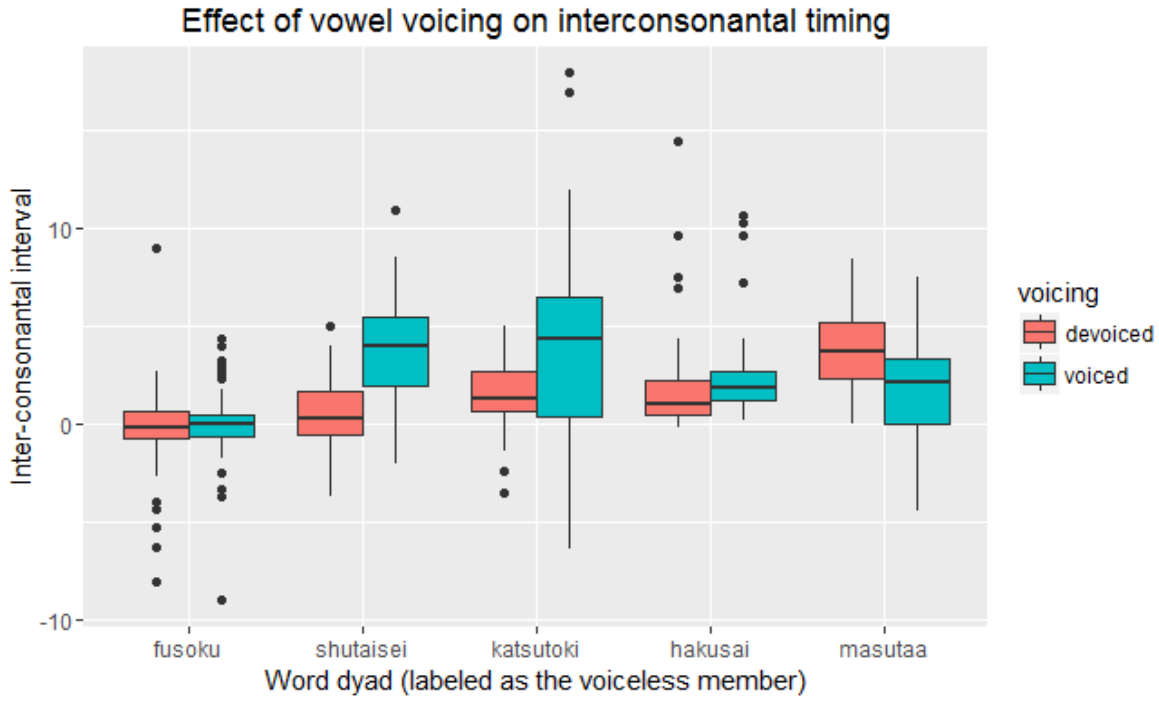


Figure 9: Inter-consonantal interval duration for each item, classified by target absence/presence.

Table 4: Model comparison showing effect of (de)voicing on ICI duration.

Model of ICI	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
DYAD + (1|speaker)	7	7869	7901	-3928	-	-
DYAD + VOWEL_VOICING + (1|speaker)	8	7851	7887	-3917	20.39	0.000006***
DYAD* VOWEL_VOICING + (1|speaker)	12	7802	7856	-3889	56.99	1.2e-11***

In the light of our analysis of vowel height targets in the preceding section, it is notable that the two dyads that showed the greatest effect of devoicing on ICI—*ʃutaise:~ʃudaika* and *katsutoki~katsudo*—are also those that have the highest probability of vowel height targetlessness. Given this pattern of results, it is possible that the effect of vowel voicing on ICI may be attributable to the occasional (height) targetlessness of flanking vowels. The intuitive idea is that consonants move closer together when the intervening vowel is reduced or absent. We assessed this possibility through further model comparison. In one comparison, we replaced the VOWEL VOICING factor with another factor, TARGET PRESENCE, which we determined according to the Bayesian classification of tongue dorsum trajectories. We also fit a model including both factors, VOWEL VOICING and TARGET PRESENCE, to see if TARGET PRESENCE would explain variance in ICI above and beyond the VOWEL VOICING factor. We note here that these model comparisons required that we exclude the *hakusai~jakuzai* dyad (see discussion in previous section) because of the ambiguity that the velar consonant introduces into the classification results. We therefore refit the baseline models to the subset of data for which we were able to report classification results (525 tokens). As with the models of the full data set reported in Table 4 (665 tokens), including an interaction between DYAD and VOWEL VOICING showed significant improvement over baseline. However, as shown in Table 5, adding TARGET PRESENCE and the interaction between TARGET PRESENCE and DYAD as fixed factors, despite the

added complexity, resulted in only marginal improvement (p = .07). Moreover, the direct comparison of VOWEL VOICING and TARGET PRESENCE rendered negligible differences. From these results, we conclude that the observed effect of devoicing on ICI is not due to vowel height targetlessness alone; rather, it appears to be a genuine effect of devoicing. This is true despite its largest effects coming in words that also happen to frequently lack vowel height targets.

Table 5: Model comparison showing effect of vowel height target on ICI duration.

Model of ICI	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
DYAD*VOWEL_VOICING + (1|speaker)	10	6194	6236	-3086		
DYAD*VOWEL_VOICING + DYAD* TARGET_PRESENSE + (1|speaker)	14	6193	6253	-3083	8.67	0.0698

To summarize the results on the ICI interval, we found that the effects of vowel devoicing show substantial variation across words. Some combinations of oral gestures are pulled closer together in time when they flank a devoiced vowel than when they flank a voiced vowel. However, shortening of ICI does not seem to be due to the occasional absence of a vowel height target. Rather, the chain of causation may run the other direction. Vowel reduction, including the absence of a height target, may be driven in part by the proximity of the flanking consonants. Supra-laryngeal gestures brought closer together in time due to laryngeal reorganization may encourage non-specification of vowel gestures on one or many dimensions.

We conclude the presentation of the results by evaluating the correlation between ICI duration and C_1 duration. The predicted relation between these intervals depends on coordination topology (Table 1). As long as the /u/ is present (phonetically specified), we assume that it will

be coordinated with the preceding consonant, i.e., C-V coordination (Browman & Goldstein, 2000; Gafos, 2002; Smith, 1995). If /u/ is absent, then the consonants flanking the vowel may be timed to each other, i.e., C-C coordination (Browman & Goldstein, 2000; Gafos, 2002; Shaw & Gafos, 2015). As summarized in Table 1, C-V coordination predicts a trade-off (negative correlation) between C_1 duration and ICI duration in our data, because as C_1 shortens, it will expose more of the vowel and ICI will lengthen. C-C coordination on the other hand, predicts no such relation between C_1 and ICI. However, all else equal, we expect a positive correlation between adjacent intervals, such as C1 and ICI, as both would be influenced by common factors, such as speech rate and ease of lexical access. Investigating the correlation between C_1 and ICI thus affords the opportunity for an assessment of vowel presence independent from our analysis of TD trajectories, which focused only on the height dimension. A negative correlation between C_1 and ICI, a characteristic of C-V coordination, provides evidence, albeit indirect, that a vowel gesture is present in the signal. Figure 10 shows a scatterplot of C_1 duration and ICI. Since we are comparing across items that have different average ICI, we have z-scored C1 duration and ICI within item to avoid obtaining spurious correlation. The blue triangles are tokens that contain a vowel height target, according to our analysis of TD trajectories. The red triangles represent tokens that lack a height target. The blue and red lines are linear fits to the tokens with and without vowel height targets, respectively.

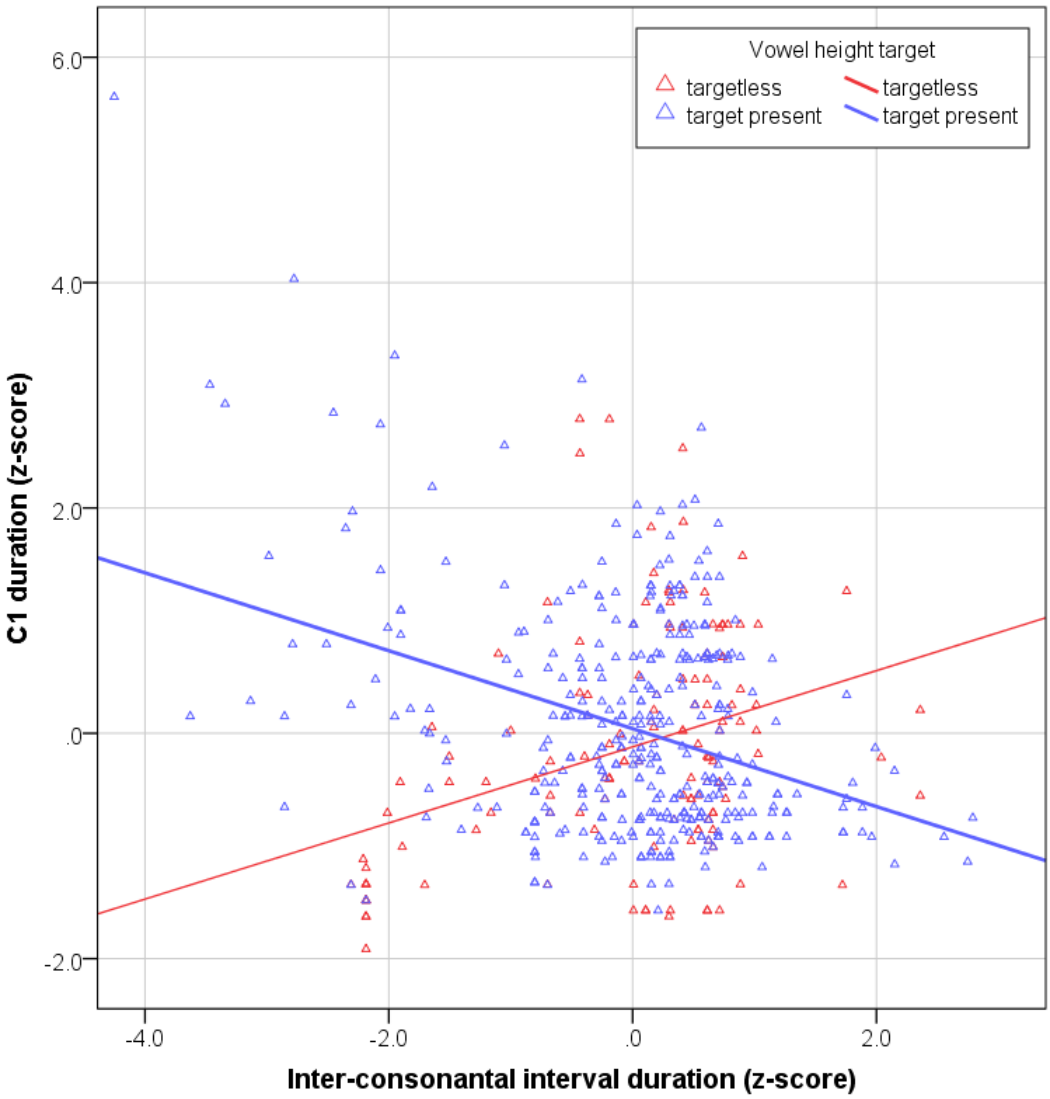


Figure 10: The correlation between C1 duration (y-axis) and ICI (x-axis) across all items in the corpus. Blue triangles represent voiced vowels and devoiced vowels classified as having a vowel height target. Red triangles represent devoiced vowels classified as lacking a vowel height target. The linear regression line fit to the blue triangles (target present cases) shows a significant negative trend; the red triangles (target absent cases) show the opposite direction of correlation.

The distributions of C1 and ICI values both showed significant deviations from normality (C1 was right skewed; ICI was left skewed) according to a Shapiro-Wilkes test ($C1_{target\_present}$: $w(386) = .894$, $p < .001$; $C2_{targetless}$: $w(139) = .895$, $p < .001$; $ICI_{target\_present}$: $w(386) = .975$, $p < .001$; $ICI_{targetless}$: $w(139) = .969$, $p = .003$). We therefore conducted nonparametric Spearman correlation analyses. There was a significant negative correlation between C1 and ICI when the vowel height target is present ($\rho(386) = -0.19$; $p < .001$). When absent, the relation between C1 and ICI is positive and also statistically significant ($\rho(139) = 0.289$, $p < .001$). Thus, the negative tradeoff between consonant duration and ICI duration is only maintained when the lingual vowel height target is present. When the vowel is absent, C1 is positively correlated with ICI. This provides a converging argument for the variation in vowel specification found across tokens in our analysis of TD height trajectories. There are clearly systematic differences in temporal organization between those tokens classified as containing a vowel height target and those tokens classified as lacking a vowel height target. Moreover, the direction of the differences is as expected is C-V coordination were only available for tokens that contain a vowel height target. The absence of the negative correlation between C1 and ICI for tokens that lack a vowel height target suggests that at least some of these tokens lack sufficient specification to enter into C-V coordination.

The analyses of tongue dorsum trajectories and of C-V timing provide converging support for H4, the hypothesis that devoiced vowels in Japanese are optionally targetless. Change over time in tongue dorsum height, the most salient dimension of Japanese /u/, approximated a linear trajectory between flanking non-high vowels in some devoiced tokens. In these tokens, the tongue dorsum did not rise from its position for the preceding non-high vowel to /u/. Linear interpolation of TD height across VCuCV sequences alternated with another pattern. Some tokens containing devoiced /u/ were indistinguishable from voiced /u/—that is, the tongue

dorsum rose from its position for the preceding non-high vowel to achieve a vowel height target for /u/ before lowering again for the following non-high vowel.

Speakers differ in the degree to which they produced words without a vowel height target for devoiced /u/, but all speakers produced /u/ without a lingual target in some words some of the time. When the vowel height target was present (but not when it was absent), we observed a negative correlation between C_1 duration and ICI, as predicted by C-V coordination (Table 1). The absence of this correlation across tokens lacking a vowel height target suggests a different coordination regime, i.e., C-C coordination. The tendency to produce a targetless vowel also varies across words, possibly due to the consonantal environment in which /u/ occurs. Devoicing also impacted the timing between flanking consonants. The interval between consonants flanking /u/ (ICI) decreased with devoicing, a pattern which also varied significantly across words. Notably, the effect of devoicing on ICI was greatest for /ʃutaise:/ and /katsutoki/, the words which, on average, showed the highest degree of height targetlessness.

Discussion

The primary purpose of the study was to examine the lingual articulation of devoiced vowels in Japanese. One of our foci was the height dimension of /u/. We focused on /u/, as opposed to /i/ (which also devoices in Tokyo Japanese), since there is no previous lingual articulatory data on this vowel. We focused on the height dimension because height is the most salient characteristic of Japanese /u/. Even for voiced variants, the degree of backness and rounding in Tokyo Japanese /u/ is often reduced relative to how /u/ is defined by the IPA. At the outset, we formulated four specific hypotheses based on previous research, which we tested in an EMA experiment. The data were analyzed via Bayesian classification of DCT components fit to the TD trajectories. One innovative aspect of our analysis is that we defined a category lacking a

vowel height target in terms of a linear trajectory between flanking vowels. This approach allowed us to consider on a token-by-token basis whether the vowel was specified for a height target. Results support the hypothesis that vowel height targets are optional in devoiced vowels. When the devoiced vowels in our study were produced without a height target, the TD height trajectory for /u/ followed a path of linear interpolation between flanking vowels. All speakers produced at least some tokens that were classified as linear interpolations, although the probability of such tokens varied across speakers and across words.

Recall that there are competing claims in past work about the environments that condition devoicing vs. deletion (Kawakami 1971; Whang 2014). Comparing Table 2, which shows the claims of past work, and Table 3, which shows the actual deletion probabilities, our results do not match with either. In particular, both Kawakami (1971) and Whang (2014) argue for devoicing, not deletion, in [ʃutaise:], but we found the highest probability of height targetlessness in that environment. Across speakers, vowel height targets were identified most often in /masutaa/ followed by /ɸu̲soku/ and then /katsutoki/ and /ʃu̲taise:/. These four items differ in many ways which may contribute to this result, including lexical statistics, the presence/absence of morphological boundaries, and aspects of the vowel and consonantal environments of the target /u/. For example, although the target /u/ was always flanked by non-high vowels, the precise vowel contexts were different across target items: /a/-/u/-/a/ in /masuta/, /e/-/u/-/o/ in /ɸu̲soku/, /a/-/u/-/o/ in /katsutoki/, and /e/-/u/-/a/ in /ʃu̲taise:/. We note that the /a/-/u/-/a/ in /masutaa/ requires the largest articulatory movement for /u/, and it is here where we observe vowel height targets most frequently. It could be that the lingual articulatory difference between voiced and voiceless vowels are conditioned by coarticulatory context such that differences between voiced and devoiced vowels are minimized for more extreme

articulatory trajectories.[10] It is also the case that /masutaa/ is the only recent loanword in the study and that it derives from a word in English with a /st/ cluster, although this etymology did not seem to encourage vowel target absence. We also note that the /e/ preceding target /u/ in /e#ɸu̲soku/ and /e#ʃu̲taise:/ comes from the carrier phrase, while the preceding vowels in the other items are internal to the target word. While we took steps to discourage insertion of a phonological phrase boundary between the carrier phrase and the target word (see methods), as this is known to influence vowel-to-vowel coarticulation (e.g., Cho, 2006), the presence of a morpheme boundary may also affect timing across gestures (Cho, 2001; Lee-Kim, Davidson, & Hwang, 2013)), leading to differences across items. Specifically, increased vowel overlap across morphological boundaries could reduce the rise in TD trajectory for both voiced and voiceless /u/, pushing both towards the linear interpolation trajectory that served as the basis for our "target absent" classification. We also not that, as there is also a morpheme boundary in /katsu#toki/, the three items with the greatest incidence of height targetlessness contained a morpheme boundary within the target VCVCV trajectory. The flanking consonants may also play a role in conditioning item specific differences. The degree to which vowels coarticulate across consonants is known to be influenced by the degree of palatal contact such that a consonant like /ʃ/ with a high degree of palate contact has a greater degree of coarticulatory resistance than /t/ which has less contact with the palate (Recasens, 1989). Although consonantal context was controlled across voiced-voiceless dyads, this could potentially be a source of difference across dyads. Our analysis of the inter-consonantal interval (ICI) showed that some combinations of consonants, /ʃ_t/ and /ts_t/ in particular, were pulled closer together when intervening vowels are devoiced, regardless of targetlessness. It was in these contexts that

[10] We would like to thank the editor, Taehong Cho, for this suggestion.

we also observed increased frequency of targetlessness. This suggests a possible chain of causation whereby laryngeal reorganization dictated by devoicing pulls consonants closer together which obliterates the intervening vowel. Although teasing apart the various factors that may be conditioning item difference requires future studies, we find the role of flanking consonants to be one of the most intriguing possibilities.

We acknowledge that our focus on height in the analysis of TD trajectories prevents us from drawing any firm conclusions about complete vowel absence from the classification analysis alone. As mentioned above, height is the most salient feature of /u/ in Japanese. Nevertheless, linear interpolation on the height dimension, indicating a lack of vowel height target, does not preclude phonetic specification in other dimensions. Past articulatory data from ultrasound and MRI indicate that the highest position of the tongue for /u/ is rather central, c.f., the substantially more posterior position observed for /o/; the labial component of /u/ has long been recognized as different from similar instances of this phone in other languages (Vance, 2008). Neither the labial nor backing components of /u/ involve particularly large movements, making it difficult to discern differences between voiced and devoiced /u/.

Besides the possibility of lip compression or backness targets for /u/, there is, as mentioned in the introduction, evidence that laryngeal gestures originating with voiceless consonants are controlled, c.f., passive coarticulation, to yield vowel devoicing. This conclusion goes back at least to Sawashima (1971),[11] who writes (pg 13): “The opening of the glottis for the medial [kt] and [kk] were significantly larger than for [tt] and [kk] which lasted for approximately the same

[11] We would like to thank an anonymous reviewer for pointing us to this literature.

durations, and this fact shows that the glottal adjustments for devoicing of the vowel are not a mere skipping of the phonatory adjustments for the vowel but a positive effort of widening of the glottis for the devoiced vowel segment, even though there is no phonemic distinction between the voiced and devoiced vowels." Moreover, there is evidence from EPG that the tongue maintains fricative-like contact with the palate during devoiced vowels, which may be controlled to sustain devoicing. These aspects of devoiced vowel articulation likely persist (we have no evidence to indicate that they don't) even when the tongue does not move towards a height target for /u/. For these reasons, we cannot equate the absence of a height target with the absence of vowel. Nevertheless, the data strongly suggest that the height target is categorically present or absent across tokens, even if other aspects of the devoiced vowel remain under speaker control.

One theoretical consequence of our results is that they constitute a categorical but optional pattern. As recently pointed out by Bayles, Kaplan & Kaplan (2016), "optionality" in many studies can come from averaging over inter-speaker differences, and it is important to examine whether a categorical pattern can show true optionality *within a speaker*. In the case of devoiced vowels in Japanese, the average trajectory of devoiced /u/ would point to the misleading conclusion that /u/ is reduced, because it averages over "full target" and "no target" tokens (see Shaw & Kawahara, submitted-a for further details). Although this was not the main purpose of this experiment, our finding supports the view expressed by Bayles et al. (2016) that indeed, a categorical pattern, like French schwa deletion, can be optional within a speaker. Other articulatory research has identified similarly optional patterns, including nasal place assimilation in English (Ellis & Hardcastle, 2002) and place assimilation in Korean (Kochetov & Pouplier, 2008). Revealing such patterns requires that the phonological status of each token is assessed individually. The data from Bayles et al. (2016) draws from an already segmented

corpus, essentially trusting the judgements about when a vowel appears or does not appear. Ellis & Hardcastle (2002) identified optional patterns of nasal place assimilation in English through visual inspection of EMA and EPG data. They collected baseline data of a (lexically) velar nasal followed by a velar stop to evaluate possible assimilation of a coronal nasal to a following velar stop. Kochetov & Pouplier (2008) went beyond visual inspection but found similar patterns. For their study of place assimilation in Korean, they established a criterion—two standard deviations from the mean of the canonical category—which they used to classify tokens as either fully articulated (within two standard deviations) or reduced. Our approach goes one step further, as we explicitly define categories based both on the full target and on phonetic interpolation (target absent scenario). Despite some differences in analytical method, we found a similar type of variation as Bayles et al. (2016), i.e., within-item and, in many cases, within speaker variation of a largely categorical nature. To these results we can add vowel height specification (presence/absence) to the list of categorical but optional processes.

There are numerous theoretical frameworks capable of handling the optional presence/absence of a vowel or even a particular vowel feature. These include varbrul (Guy, 1988) and exemplar theory (Pierrehumbert, 2006) as well as generative models developed to handle categorical variation, including Stochastic OT (Hayes & Londe, 2006) and (Noisy) Harmonic Grammar (Coetzee & Kawahara, 2013; McPherson & Hayes, 2016). In contrast to other optional phonological patterns to which these models have been applied, the Japanese vowel devoicing case is of particular theoretical interest because of learnability issues. Most theoretical approaches to variability proceed by matching frequencies in the input data. Devoicing impoverishes the acoustic signature of tongue dorsum height, disrupting as well the auditory feedback that the learner may receive from variation in their own production of /u/. Hence, the degree to which a learner can match productions in their own speech to frequencies in the

ambient environment is limited in this case. Learners may be restricted to somatosensory feedback from their own productions (Tremblay, Shiller, & Ostry, 2003). Relevant to the learnability issue is the observation that infant-directed speech in Japanese contains vowel devoicing at approximately the same rates as adult-directed speech (Fais, Kajikawa, Amano, & Werker, 2010). To the extent that systematic patterns, e.g., conditioning environments for vowel height targetlessness, are found in the population, they may emerge from analytical bias as opposed to pattern matching (Moreton, 2008).

The continuity of phonetic measurements makes it natural to consider the possibility that a gesture is categorically present but reduced in magnitude or coordinated in time with other gestures such that the acoustic consequences of the gesture are attenuated (Browman & Goldstein, 1992b; Iskarous, McDonough, & Whalen, 2012; Jun, 1996; Jun & Beckman, 1993; Parrell & Narayanan, 2014). Particularly with vowels, gradient and gradual phonetic shift is well-documented and is often treated as the primary mechanism of variation (e.g., Labov, Ash, & Boberg, 2005; Wells, 1982 for comprehensive overviews). This underscores the importance of deploying rigorous methods to support claims about the presence/absence of a gesture in the phonetic signal of which the Japanese data offer a clear case. We emphasize at this point also, that our method is useful in testing the general "phonetic underspecification analysis" (Keating, 1988). Several studies have argued that certain segments lack phonetic targets in some dimensions (Cohn, 1993; Keating, 1988; Pierrehumbert & Beckman, 1988). By rigorously relating the signal to both a full vowel and the linear interpolation hypothesis, the current computational analysis offers a general approach that can be used to evaluate phonetic underspecification.

The variable targetlessness of devoiced /u/ raises several new research questions. Firstly, we have observed that targetless probabilities differ across words, but we have left the precise conditioning environments to future study. The identity of the flanking consonants, the lexical statistics of the target words, the informativity of the vowel, etc., are all possibilities. Due to the small number of words recorded in this experiment we hesitate to speculate on which of these factors (and to what extent) may influence targetlessness, but we plan to follow up with another study that expands the number of words instantiating the consonantal environments reported here. A second question is the syllabic status of consonant clusters that result from targetless /u/. Matsui (2014) speculates that the preceding consonant forms an independent syllable, a syllabic consonant, while Kondo (1997) argues that it is resyllabified into the following syllable, forming a complex onset. A third question is the status of /i/, the other of the two high vowels that are categorically devoiced in Tokyo Japanese. Our conclusion thus far is solely about /u/, which is independently known to be variable in its duration (Kawahara & Shaw, submitted-b), and may be more susceptible to coarticulation than /i/ (c.f., Recasens & Espinosa, 2009). At this point we have nothing to say about whether devoiced /i/ may also be targetless in some contexts but hold this to be an interesting question for future research. Finally, the observed shift from C-V to C-C coordination may bear on the broader theoretical issue of how higher level structure, i.e., the mora in Japanese, relates to the temporal organization of consonant and vowel gestures. The CV mora may be less determinate of rhythmic patterns of Japanese than is sometimes assumed (Beckman 1982; Warner and Arai, 2001).

Conclusion

The current experiment was designed to address the question of whether devoiced /u/ in Japanese is simply devoiced (Jun & Beckman 1993) or whether it is also targetless (Kondo

2001). Since previous studies on this topic have used acoustic data (Whang 2014) or impressionistic observations (Kawakami 1977), we approached this issue by collecting articulatory data. Using EMA, we recorded tongue dorsum trajectories in words containing voiced and devoiced /u/. Some devoiced tokens showed a linear trajectory between flanking vowels, indicating that there is no height target for the vowel. Other devoiced vowel tokens had trajectories like voiced vowels. Tokens that were reduced, showing trajectories intermediate between the voiced vowels and linear interpolation between flanking vowels, were less common. We conclude that /u/ is optionally targetless; i.e., there is token-by-token variability in whether the lingual gesture is present or absent. The patterns of covariation between C1 duration and ICI provided further support for this conclusion. For tokens classified as containing a lingual target, there is a significant negative correlation between C1 duration and ICI, a prediction of C-V timing. The correlation is in the opposite direction for tokens that lacks a lingual height target for the vowel.

Achieving the above descriptive generalization—that devoiced vowels are optionally targetless—required some methodological advancements, including analytical tools for assessing phonological status on a token-by-token basis. We analyzed the data using Bayesian classification of a compressed representation of the signal based on Discrete Cosine Transform (following Shaw & Kawahara, submitted-a). The posterior probabilities of the classification showed a bimodal distribution, supporting the conclusion that devoiced /u/ in Tokyo Japanese is variably targetlessness.

Overall, establishing that devoiced /u/ tokens are sometimes targetless, in that they do not differ from the linear interpolation of flanking gestures, answers a long-standing question in Japanese phonetics/phonology while raising several new questions to pursue in future research, including

the syllabic status of consonant clusters flanking a targetless vowel, the role of the mora (or lack thereof) in Japanese timing, the phonological contexts that favor targetless /u/, and whether other devoiced vowels, particularly /i/, may also be variably targetless.

Acknowledgements

This project is supported by JSPS grant #15F15715 to the first and second authors, #26770147 and #26284059 to the second author. Thanks to audience at Yale, ICU, Keio, RIKEN, and Phonological Association in Kansai and "Syllables and Prosody" workshop at NINJAL, in particular Mary Beckman, Lisa Davidson, Junko Ito, Michinao Matsui, Reiko Mazuka, Armin Mester, Haruo Kubozono, and three anonymous reviewers. All remaining errors are ours.

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*(1), 31-56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of Acoustical Society of America, 119*(5), 3048-3059.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version, 1*(7).

Bayles, A., Kaplan, A., & Kaplan, A. (2016). Inter-and intra-speaker variation in French schwa. *Glossa: a journal of general linguistics, 1*(1).

Beckman, M. (1982). Segmental duration and the 'mora' in Japanese. *Phonetica, 39*, 113-135.

Beckman, M. (1986). *Stress and Non-Stress Accent.* Dordrecht: Foris.

Beckman, M., & Shoji, A. (1984). Spectral and Perceptual Evidence for CV Coarticulation in Devoiced/si/and/syu/in Japanese. *Phonetica, 41*(2), 61-71.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92-111.

Berry, J. J. (2011). Accuracy of the NDI wave speech research system. *Journal of Speech, Language, and Hearing Research, 54*(5), 1295-1301.

Blackwood-Ximenes, A., Shaw, J., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel variation across American and Australian dialects of English. *The Journal of Acoustical Society of America, 142*(2), 363-377.

Bombien, L., Mooshammer, C., & Hoole, P. (2013). Articulatory coordination in word-initial clusters of German. *Journal of Phonetics, 41*(6), 546-561.

Browman, & Goldstein, L. (1992a). 'Targetless' schwa: An articulatory analysis. In G. Docherty & R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 26-56). Cambridge: Cambridge University Press.

Browman, C., & Goldstein, L. (1992b). Articulatory phonology: An overview. *Phonetica, 49*, 155-180.

Browman, C. P., & Goldstein, L. M. (2000). Competing Constraints on Intergestural Coordination and Self-Organization of Phonological Structures. *Les cahiers de l'lCP, Bulletin de la Communication Parlee, 5*, 25-34.

Byrd, D. (1996). Influences on Articulatory Timing in Consonant Sequences. *Journal of Phonetics, 24*(2), 209-244.

Cedergren, H. J., & Simoneau, L. (1985). La chute des voyelles hautes en français de Montréal:'As-tu entendu la belle syncope?'. *Les tendances dynamiques du français parlé à Montréal, 1*, 57-145.

Chitoran, I., Goldstein, L., & Byrd, D. (2002). Gestural overlap and recoverability: articulatory evidence from Georgian. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 419-447). Berlin, New York: Mouton de Gruyter.

Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica, 58*(3), 129-162.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics, 32*(2), 141-176.

Coetzee, A. W., & Kawahara, S. (2013). Frequency biases in phonological variation. *Natural Language & Linguistic Theory, 31*(1), 47-89.

Coetzee, A. W., & Pater, J. (2011). The place of variation in phonological theory. *The Handbook of Phonological Theory, Second Edition*, 401-434.

Cohn, A. C. (1993). Nasalisation in English: phonology or phonetics. *Phonology, 10*(01), 43-81.

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

Davis, C., Shaw, J., Proctor, M., Derrick, D., Sherwood, S., & Kim, J. (2015). Examining speech production using masked priming. *18th International Congress of Phonetic Sciences.*

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1568-1578.

Eftychiou, E. (2010). Routes to lenition: an acoustic study. *PLoS One, 5*(3), e9828.

Ellis, L., & Hardcastle, W. J. (2002). Categorical and gradient properties of assimilation in alveolar to velar sequences: evidence from EPG and EMA data. *Journal of Phonetics, 30*(3), 373-396.

Faber, A., & Vance, T. J. (2000). More acoustic traces of 'deleted'vowels in Japanese. *Japanese/Korean Linguistics, 9*, 100-113.

Fais, L., Kajikawa, S., Amano, S., & Werker, J. F. (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, M. (2015). Chapter 4: Vowel devoicing. In H. Kubozono (Ed.), *The handbook of Japanese phonetics and phonology.* Berlin: Mouton de Gruyter.

Fujimoto, M., Murano, E., Niimi, S., & Kiritani, S. (2002). Differences in glottal opening pattern between Tokyo and Osaka dialect speakers: factors contributing to vowel devoicing. *Folia phoniatrica et logopaedica, 54*(3), 133-143.

Gafos, A. (2002). A grammar of gestural coordination. *Natural Language and Linguistic Theory, 20*, 269-337.

Gafos, A., & Goldstein, L. (2012). Articulatory representation and organization. *The handbook of laboratory phonology*, 220-231.

Gafos, A., Hoole, P., Roon, K., & Zeroual, C. (2010). Variation in timing and phonological grammar in Moroccan Arabic clusters. In C. Fougeron, B. Kühnert, M. d'Imperio, & N. Vallée (Eds.), *Laboratory Phonology* (Vol. 10, pp. 657-698). Berlin, New York: Mouton de Gruyter.

Gafos, A., Kirov, C., & Shaw, J. (2010). *Guidelines for using Mview*.

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis, 54*(4), 1167-1178.

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. *Frontiers in phonetics and speech science*, 239-250.

Guy, G. (1988). Advanced VARBRUL analysis. *Linguistic change and contact*, 124-136.

Hall, K. C., Hume, E., Jaeger, F., & Wedel, A. (2016). *The Message Shapes Phonology*.

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation,/u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the acoustical society of America, 123*(5), 2825-2835.

Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology, 23*(1), 59-104.

Hirayama, M. (2009). Postlexical Prosodic Structure and Vowel Devoicing in Japanese (2009). *Toronto Working Papers in Linguistics*.

Hirose, H. (1971). The activity of the adductor laryngeal muscles in respect to vowel devoicing in Japanese. *Phonetica, 23*(3), 156-170.

Imaizumi, S., & Hayashi, A. (1995). Listener-adaptive adjustments in speech production: evidence from vowel devoicing. *Ann. Bull. RIPL, 29*, 43-48.

Iskarous, K., McDonough, J., & Whalen, D. (2012). A gestural account of the velar fricative in Navajo.

Isomura, K. (2009). *Nihongo-wo Oshieru [Teaching Japanese]*. Tokyo Hitsuji.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of Acoustical Society of America, 94*(2), 701-714.

Jun, J. (1996). Place assimilation is not the result of gestural overlap: evidence from Korean and English. *Phonology, 13*(03), 377-407.

Jun, S.-A., & Beckman, M. (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, S.-A., Beckman, M. E., & Lee, H.-J. (1998). Fiberscopic evidence for the influence on vowel devoicing of the glottal configurations for Korean obstruents. *UCLA Working Papers in Phonetics*, 43-68.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language, 45*, 229-254.

Kawahara, S. (2015). A catalogue of phonological opacity in Japanese: Version 1.2. *慶応義塾大学言語文化研究所紀要*(46), 145-174.

Kawakami, S. (1977). Outline of Japanese Phonetics [written in Japanese as" Nihongo Onsei Gaisetsu"]: Tokyo: Oofuu-sha.

Keating, P. (1988). Underspecification in phonetics. *Phonology, 5*, 275-292.

Kilbourn-Ceron, O., & Sonderegger, M. (2017). Boundary phenomena and variability in Japanese high vowel devoicing. *Natural Language & Linguistic Theory*, 1-43.

Kingston, J. (1990). Articulatory binding. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I* (pp. 406-434). Cambridge: Cambridge University Press.

Kochetov, A., & Kang, Y. (2017). Supralaryngeal implementation of length and laryngeal contrasts in Japanese and Korean. *Canadian Journal of Linguistics/Revue canadienne de linguistique, 62*(1), 18-55.

Kochetov, A., & Pouplier, M. (2008). Phonetic variability and grammatical knowledge: An articulatory study of Korean place assimilation. *Phonology, 25*(3), 399-431.

Kondo, M. (1997). Mechanisms of vowel devoicing in Japanese.

Kondo, M. (2001). Vowel Devoicing and Syllable Structure in Japanese. In M. Nakayama & C. J. Quinn (Eds.), *Japanse/Korean Linguistics* (Vol. 9). Stanford: CSIL.

Kondo, M. (2005). Syllable structure and its acoustic effects on vowels in devoicing environments. *Voicing in Japanese, 84*, 229.

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

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*: Walter de Gruyter.

Lee-Kim, S.-I., Davidson, L., & Hwang, S. (2013). Morphological effects on the darkness of English intervocalic/l. *Laboratory Phonology, 4*(2), 475-511.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory *Speech production and speech modelling* (pp. 403-439): Springer.

Maekawa, K. (1990). *Production and perception of the accent in the consecutively devoiced syllables in Tokyo Japanese.* Paper presented at the ICSLP.

Maekawa, K., & Kikuchi, H. (2005). Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In J. v. d. Weijer, K. Nanjo, & T. Nishihara (Eds.), *Voicing in Japanese* (pp. 205-228). Berlin, New York: Mouton de Gruyter.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., . . . Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language resources and evaluation, 48*(2), 345.

Marin, S. (2013). The temporal organization of complex onsets and codas in Romanian: A gestural approach. *Journal of Phonetics, 41*(3), 211-227.

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gesture coupling model. *Motor Control, 14*, 380-407.

Matsui, M. (2014). Vowel devoicing, VOT Distribution and Geminate Insertion of Sibilants 歯擦音の母音無声化・VOT 分布・促音挿入. *Theoretical and applied linguistics at Kobe Shoin: トークス, 17*, 67-106.

McPherson, L., & Hayes, B. (2016). Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology, 33*(1), 125-167.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology, 25*(01), 83-127.

Munhall, K., & Lofqvist, A. (1992). Gestural aggregation in speech: laryngeal gestures. *Journal of Phonetics, 20*(1), 111-126.

Nakamura, M. (2003, August 3-9). *The Spatio-temporal Effects of Vowel Devoicing on Gestural Coordination: an EPG study.* Paper presented at the 15th International Congress of Phonetics Sciences, Barcelona.

Nielsen, K. Y. (2015). Continuous versus categorical aspects of Japanese consecutive devoicing. *Journal of Phonetics, 52*, 70-88.

Nogita, A., Yamane, N., & Bird, S. (2013). *The Japanese unrounded back vowel/ɯ/is in fact rounded central/front [ʉ-ʏ]*. Paper presented at the Ultrafest VI, Edinburgh.

Parrell, B., & Narayanan, S. (2014). *Interaction between general prosodic factors and languagespecific articulatory patterns underlies divergent outcomes of coronal stop reduction.* Paper presented at the International Seminar on Speech Production (ISSP) Cologne, Germany.

Pierrehumbert, J. (2006). The next toolkit. *Journal of Phonetics, 34*(6), 516-530.

Pierrehumbert, J., & Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, Mass.: MIT Press.

Pierrehumbert, J. B. (2001). Stochastic phonology. *Glot International, 5*(6), 195-207.

Poser, W. J. (1990). Evidence for foot structure in Japanese. *Language, 66*, 78-105.

Recasens, D. (1989). Long range coarticulation effects for tongue dorsum contact in VCVCV sequences. *Speech Communication, 8*(4), 293-307.

Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the acoustical society of America, 125*(4), 2288-2298.

Sawashima, M. (1971). Devoicing of vowels. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, 5*, 7-13.

Shaw, J. A., Chen, W.-r., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research, 59*(6), S1566-S1574.

Shaw, J. A., Gafos, A., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology, 28*(3), 455-490.

Shaw, J. A., & Gafos, A. I. (2015). Stochastic Time Models of Syllable Structure. *PLoS One, 10*(5), e0124714.

Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology, 26*, 187-215.

Shaw, J. A., & Kawahara, S. (submitted-a). A computational toolkit for assessing phonological specification in phonetic data: Discrete Cosine Transform, Micro-Prosodic Sampling, Bayesian Classification. *Phonology*, 34.

Shaw, J. A., & Kawahara, S. (submitted-b). Effects of entropy and suprisal on vowel duration in Japanese. *Language and Speech*.

Sjoberg, A. F. (1963). *Uzbek structural grammar* (Vol. 18): Indiana university.

Smith, C. L. (1995). Prosodic patterns in the coordination of consonant and vowel gestures. In B. Connell & A. Arvaniti (Eds.), *Papers in laboratory phonology IV: phonology and phonetic evidence* (pp. 205-222). Cambridge: Cambridge University Press.

Smith, C. L. (2003). Vowel devoicing in contemporary French. *Journal of French Language Studies, 13*(02), 177-194.

Sugito, M., & Hirose, H. (1988). Production and perception of accented devoiced vowels in Japanese. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics, 22*, 19-36.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*(6942), 866.

Tsuchida, A. (1997). *Phonetics and phonology of Japanese vowel devoicing*. Ph. D. Dissertation: University of Cornell.
Vance, T. (1987). *An Introduction to Japanese Phonology*. New York: SUNY Press.
Vance, T. J. (2008). *The sounds of Japanese with audio CD*: Cambridge University Press.
Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica, 58*(1-2), 1-25.
Wells, J. C. (1982). *Accents of English, vol.2: The British Isles*. Cambridge: Cambridge University Press.
Whang, J. (2014). Effects of predictability on vowel reduction. *Journal of Acoustical Society of America, 135*(4), 2293.
Wright, R. (1996). *Consonant Clusters and Cue Preservation in Tsou.* (PhD dissertation), UCLA, Los Angeles.
Yip, J. C.-K. (2013). *Phonetic effects on the timing of gestural coordination in Modern Greek consonant clusters.* University of Michigan.

Appendix

The number of consonant gestures by condition parsed by tangential vs. component velocities.

		C1		C2	
		voiceless	voiced	voiceless	voiced
ɸusoku~ ɸuzoku	tangential velocity	68	70	63	66
	component velocity	1	1	6	5
ʃutaise:~ʃudaika	tangential velocity	31	42	50	24
	component velocity	36	28	20	43
katsutoki~katsudo:	tangential velocity	50	55	50	50
	component velocity	23	11	20	19
hakusai~yakuzai	tangential velocity	50	55	70	67
	component velocity	23	11	0	3
masutaa~masuda	tangential velocity	70	29	70	35
	component velocity	0	10	0	4