To appear in LabPhon

Limits on gestural reorganization following vowel deletion: The case of Tokyo Japanese

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

Running head:gestural reorganization

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

Keywords: Japanese, devoicing, EMA, articulation, deletion, gestural coordination

Approximate word count: 10,000 words.

Abstract

The coordination of gestures in consonant clusters differs across languages and hence must be a learned aspect of linguistic knowledge. Precisely pinning down the coordination relation used in a particular language, or for a particular consonant cluster type, has been facilitated by recent research showing that coordination relations structure kinematic variation in unique ways. We apply these methods to a hitherto under-explored topic, the coordination of consonant clusters created via vowel deletion. Our case study involves fricative-fricative and fricative-stop consonant clusters resulting from the variable deletion of devoiced vowels in Tokyo Japanese. Examination of articulatory data obtained by Electromagnetic Articulography (EMA) show that some consonant clusters, i.e., fricative-stop clusters, show gestural reorganization whereas other cluster types, i.e., fricative-fricative sequences, behave as if a vowel remains in place, despite the fact that the tongue dorsum movement for the vowel is absent from the articulatory record. We discuss several theoretical possibilities to account for the differential effects of vowel deletion on gestural re-organization in these environments.

1 Introduction

1.1 General background

It has long been known that how two adjacent consonantal gestures are coordinated differs considerably across languages. Given a [tk] sequence, for example, some languages show a clear audible release for [t] whereas other languages do not. As such, such coordination patterns are a part of what speakers actively control, and hence they constitute an important part of linguistic knowledge (Gafos et al., 2020; Shaw, 2022). However, precisely pinning down the nature of coordination relations has been a difficult issue, partly because it is not always possible to infer coordination relations from impressionistic observations of speech or even from acoustic signals. The development of research methods which have allowed us to directly observe articulatory movement with high temporal resolution has made this a tractable problem. Recent work by Shaw and colleagues has demonstrated, through a number of case studies, that coordination relations between gestures can be revealed by studying the structure of temporal variation in articulatory kinematic data (e.g., Gafos et al. 2014; Shaw 2022; Shaw & Gafos 2015; Shaw & Kawahara 2018b; Shaw et al. 2021; see also Durvasula et al. 2021; Lialiou et al. 2021; Sotiropoulou & Gafos 2022). A topic that is nevertheless still under-explored is how consonant clusters created via vowel deletion are coordinated, a gap that the current paper attempts to address. Specifically, in this paper we study the coordination of consonant clusters resulting from high vowel deletion in Tokyo Japanese.

Apparent deletion of a segment can follow from a phonological process—a wholesale deletion of a phonological category—or certain patterns of gestural overlap, i.e., "gestural hiding". Extreme theoretical poles posit that all cases of apparent deletion follow from one of these sources. For example, Browman & Goldstein (1990) develop the gestural overlap hypothesis of segmental "deletion", showing how numerous cases of apparent deletion, insertion and allophony can be derived from the timing and magnitude of gestures, without necessitating symbolic transformations, including deletion. At the other end of the theoretical spectrum, allophony has been treated as transformations between linearly ordered segments. On this view, the /pət/ → [pt] mapping, as in 'potato', can only be seen as deletion, and not as gestural overlap (Chomsky & Halle, 1968; Kaisse & Shaw, 1985) (cf. Davidson 2006). By now, enough empirical evidence has been amassed to make it clear that both theoretical accounts—gestural hiding and categorical deletion—have to be retained. That is, some cases of apparent deletion, such as the /t/ in ‘perfec/t/ memory’ at fast speech are clearly present in the articulation, even though they can be masked by the overlapping lip closure (Browman & Goldstein, 1990), making them inaudible. Other cases of apparent deletion are clearly attributable to categorical deletion, even though they might plausibly have been due to overlap (Ellis & Hardcastle, 2002; Kochetov & Pouplier, 2008) (cf. Nolan 1992; see also Zsiga 2020). Studies on this topic for the last three decades have shown that without careful examination of articulatory data, it is difficult to ascertain the true source of ‘apparent’ deletions. The empirical necessity to integrate theoretical perspectives raises interesting and hitherto under-researched questions. When categorical deletion does occur, what happens to the coordination of the remaining gestures? We address this question in the current study.

To investigate this issue, it is necessary to first establish that a segment is categorically deleted using articulatory data. Only then is it possible to evaluate the coordination patterns of the resulting gestures. Tokyo Japanese presents an ideal case to investigate how vowel deletion impacts gestural coordination, because recent studies have established that devoiced vowels in this language are *variably* and *categorically* deleted (Shaw & Kawahara, 2018b, 2021). This is ideal because we can investigate coordination patterns in the same words with and without a vowel. Vowel deletion can be determined by looking at whether the tongue dorsum moves towards a target for the vowel. The timing of consonants produced with different articulators, e.g., the tongue front and the lips, can then be compared in tokens with and without a vowel, as determined by tongue dorsum movement. This is what we do in the current paper.

In the remainder of the Introduction, we summarize past work on vowel deletion in Japanese (§1.2), discuss expectations for how coordination might be impacted by vowel deletion (§1.3), and illustrate specific predictions for different coordination relations, which can be tested in kinematic data (§1.4).

1.2 Vowel deletion in Japanese

A traditional description of high vowel devoicing in Japanese is that high vowels are devoiced between two voiceless obstruents and after a voiceless obstruent word-finally. Sometimes the environment between two voiceless obstruents is further sub-divided into ‘typical’ and ‘atypical’ devoicing environments. The ‘typical’ devoicing environment is either (1) between two voiceless stops or (2) between one voiceless fricative and one voicless stop. The ‘atypical’ devoicing environment is between two voiceless fricatives (Fujimoto, 2015). Devoicing is found in both environments but it is more common (and more nearly categorical) in the ‘typical’ environments than in the ‘atypical’ environment (Maekawa & Kikuchi, 2005). There has been a long debate about the deletion status of devoiced high vowels in Japanese, with arguments that they are phonologically deleted (Beckman, 1982; Beckman & Shoji, 1984; Kondo, 2001) and also that they are merely devoiced due to overlap of the glottal abduction gestures associated with the flanking consonants (Faber & Vance, 2010; Jun & Beckman, 1993) (though see Fujimoto et al. 2002); see Fujimoto (2015) for a summary of the studies that express each point of view.

Shaw & Kawahara (2018b) contribute to this debate by conducting an experiment using EMA (Electromagnetic Articulography) and showing, in a sample of six speakers, that many tokens of devoiced [u] were produced without any tongue dorsum raising gesture, which they interpreted as vowel absence. A follow-up study replicated the result with a larger number of items and more systematic control of the surrounding consonant environment (Shaw & Kawahara, 2021).[1] In that study as well, there were numerous tokens which showed no evidence of a tongue dorsum raising gesture, and were better characterized as interpolation between surrounding vowels than as controlled movement towards a vowel target. Importantly, the vocalic gestures for /u/ in these tokens were not just simply reduced or undershot due to temporal constraints; the tongue dorsum trajectory showed a high probability of linear interpolation between flanking targets even at slow speech rates, hence supporting the categorical deletion view (Shaw & Kawahara, 2018b, 2021).[2] In this paper, we build on that result. To diagnose whether categorical deletion impacts how the resulting consonant clusters are coordinated, we analyze coordination in tokens that were classified as either having a vowel or lacking one.

1.3 Theoretical landscape: what happens to coordination when a vowel deletes?

Since there is little or no empirical data showing directly what happens to consonant coordination when a vowel deletes, we discuss possible expectations for our study based on theoretical considerations and other types of empirical data.

Perhaps the most straight-forward assumption about gestural coordination is that coordination is local (Gafos, 1999). On this assumption, the deletion of a vowel in CVC would leave the remaining two consonantal gestures, CC, locally adjacent. While the consonants may be coordinated with the vowel in CVC, they would have to be coordinated with each other in CC. On this assumption, deletion of a vowel would require a new coordination relation (i.e., gesture reorganization) because the two consonants would be coordinated with each other in CC but not in CVC (see also the schemata in Figure 2). We take this to be the standard assumption, but we also recognize that there are in fact a range of additional theoretical possibilities.

The alternative to the standard assumption would be that the coordination of gestures in CVC actually persists in CC, even in the absence of the vowel. A conceptual antecedent for this hypothesis can be found in phonological patterns. There are numerous cases in which segment deletion does not necessarily trigger additional phonological re-organization; for example, Kawahara & Shaw (2018) list a number of examples in which vowel deletion does not trigger resyllabification (see also Shaw et al. 2020). Additionally, deletion of a segment (vowel or consonant) is often incomplete in that the timing slot associated with the deleted segment persists, sometimes lengthening adjacent segments, i.e., phenomena falling under the label of "compensatory lengthening" (Kavitskaya, 2002).

Other phonological patterns have been analyzed in terms of ghost segments (Zimmermann, 2019), where a ghost segment is present for the purpose of conditioning phonological patterns but does not

---

[1] These studies only investigated patterns in the high vowel /u/ and remain agnostic about the deletion status of devoiced /i/.

[2] There is some debate on whether /u/ in Japanese is generally rounded or not (see Vance 2008), with some authors preferring to characterize the high back vowel as /ɯ/ (unrounded). The sensors on the upper and lower lip in Shaw & Kawahara (2018a) did not provide clear evidence for rounding on voiced /u/. See their supplementary materials for kinematic trajectories of the lips.

necessarily shape the phonetic signal directly. Related proposals consider gradient degrees of segment presence, which may improve the account of variable phonological patterns, such as French Liaison (Smolensky & Goldrick, 2016). Whether ghost segments, or gradiently activated segments, influence coordination is still unknown (though see Goldrick & Chu 2014 and Pouplier & Goldstein 2014 for some discussion of intra-gestural duration). On the other hand, there is some evidence that gestural coordination patterns can change even when the vowel is not deleted (e.g. Davidson 2006). This observation suggests that gestural coordination and segmental deletion may be somewhat independent.

Possibly, "deleted" vowels, i.e., vowels that lack any surface phonetic manifestation, can persist as ghost segments or ghost gestures, i.e., zero magnitude gestures, which may influence the coordination of other gestures without driving articulatory movement. Geissler (2021) raises this possibility to account for variation in gesture coordination across speakers of Diaspora Tibetan. In the sample of speakers analyzed, some had contrastive tone and others did not, but all speakers showed the coordination pattern that is characteristic of a tone gesture, i.e., all behaved as if a tonal gesture is present, even when their linguistic system lacks contrastive tones.

Besides ghost gestures, there are other theoretical hypotheses that might predict that coordination is unaffected by surface deletion of a vowel. Gestural coordination might not be strictly local. It might instead be organized according to a higher level clock, or cycle (e.g. Barbosa 2007; O'Dell & Nieminen 2019). In this case, the surface timing of consonants in CC and CVC could be identical because they stand in the same relation relative to a higher level triggering clock. For concreteness, consider a syllable-sized clock which triggers gestures according to a syllable cycle. In CVC, the first C could start at the beginning of the cycle, the V in the middle, and the second C towards the end. The consonants of CC could start at the beginning and towards the end of an abstract syllable cycle regardless of whether there is also a vowel timed to the middle of the cycle. This mechanism is no longer local, since gestures are not timed directly to each other but to an extrinsic timing mechanism.

Yet another theoretical hypothesis motivating no change in coordination following vowel deletion comes from Selection-Coordination Theory (Tilsen, 2016). In this theory, gestures that compose selection sets (which are assumed to be linguistically relevant units, such as syllables) are locally coordinated. However, gestures of different selection sets cannot be directly coordinated. This means, for example, that in a language where selection sets are syllables, vowels in adjacent syllables cannot be directly coordinated, (c.f. Smith 1995 for V-V coordination). Variable vowel deletion could be implemented in this framework as the competition between selection sets with and without a vowel, e.g., CVC vs. CC. However, if vowel deletion comes at a syllable boundary, as in CV.C → C.C, then the competition for the first selection set is between CV and C. Regardless of whether C or CV is selected in the first selection set, the gestures cannot be directly coordinated with the next C gesture, because the next C gesture is in a different selection set. Since coordination does not happen across selection sets in either case, there no real difference in coordination when the vowel is absent.

To provide a summary of the literature reviewed above, there are two broad hypotheses that emerge from theoretical and empirical considerations. Vowel deletion may or may not trigger reorganization of gestural coordination. If vowel deletion does trigger gestural reorganization, it may be the case that reorganization occurs only in certain contexts but not others.

In the strictly local coordination scenario, we expect vowel deletion in CVC (yielding CC) to result in C-C coordination, where the remaining consonants are coordinated with each other. In this case, the consonant gestures would be subject to (language-specific) constraints on C-C coordination. For example, in Moroccan Arabic, homoorganic consonant clusters have different C-C coordination than hetero-organic consonant clusters (Gafos, 2002; Gafos et al., 2010a). In Georgian, C-C coordination depends on the place of articulation of the consonants—if the first consonant is anterior to the second, there is greater overlap than if the first consonant is posterior to the second (Chitoran et al., 2002; Crouch et al., 2020). The Georgian pattern—the so-called "place-order effect"—has also been documented in other languages, particularly at faster speech rates (Gafos et al., 2010a).

As illustrated by the examples above, the nature of C-C coordination may interact with the identity of the consonants or the relation between them. It is also possible that certain consonant combinations may be more or less likely to enter into a C-C coordination relation. Cross-linguistically, fricative-stop clusters are more common across a syllable boundary than fricative-fricative clusters (Gouskova, 2004; Murray & Vennemann, 1983; Vennemann, 1988). Possibly, this is related to the relative ease of producing and perceiving these sequences (Ohala & Kawasaki-Fukumori, 1997). From this standpoint, we might expect fricative-stop clusters to reorganize to C-C coordination more readily than fricative-fricative clusters.

Specifically for Tokyo Japanese, deletion of devoiced vowels in CVC is equally likely when the vowel is flanked by two fricatives, e.g., [ɸus], as when it is flanked by a fricative and a stop, e.g., [ɸut] (Shaw & Kawahara, 2021). Moreover, in both cases, there is evidence that the initial consonant is not re-syllabified as a complex onset (Kawahara & Shaw, 2018). Rather, the initial consonant still appears to contribute a mora and syllable to the phonological representation.[3] However, it is still possible that changes in coordination are more likely for some consonant sequences than others. There may be several relevant considerations for predicting which clusters are more likely reorganize than others (cf. Gafos et al. 2020; Lialiou et al. 2021). Since, in the case at hand, we are dealing with consonants that cross a syllable boundary, e.g.,[ɸ.so] vs.[ɸ.ta] syllable contact constraints are one consideration (Gouskova, 2004; Murray & Vennemann, 1983; Vennemann, 1988). According to syllable contact laws, falling sonority, as in fricative-stop, is preferred to a sonority plateau, as in fricative-fricative, which may contribute to a tendency for fricative-stop (but not fricative-fricative) to reorganize.

Another difference between fricative-fricative and fricative-stop sequences has been found in Japanese text-setting, the process of aligning musical notes to song lyrics. Devoiced vowels between fricative-fricative consonants are more likely to be aligned to two separate musical notes than devoiced vowels between fricative-stop clusters (Starr & Shih, 2017). The devoiced vowel in FF can carry its own note, possibly because it maintains the timing of CVC instead of reorganizing. The difference in type-setting between FF and FS cannot be attributed to a difference in vowel deletion, given that there is variable vowel deletion in both environments (Shaw & Kawahara, 2021), but it might be due to a difference in how the resulting consonants are coordinated. We note as well that devoicing itself is less common in fricative-fricative contexts than in stop-fricative contexts (see discussion of 'typical' vs. 'atypical'

---

[3] Phonological evidence comes from patterns of accentuation as well as various morphophonological truncation patterns and the word minimality requirement. Kawahara & Shaw (2018) also report measures of stability indices, which support the view that these initial consonants still form their own syllables.

devoicing environments in Section 1.2), which may also be related.

To summarize, we take the standard view to be that vowel deletion triggers gesture re-organization. This is consistent with the assumption that gestural coordination is local. However, we also presented a number of theoretical reasons to expect the opposite, that coordination patterns will persist even in the absence of the vowel. Moreover, these two possible behaviors may differ according to the specific consonants involved. As motivation for this third alternative we considered several possibly related patterns in which FF and FS sequences differ. The current study aims to identify which of these three empirical possibilities is actually attested in Japanese.

1.4 Assessing changes in coordination

In order to evaluate the three possible outcomes described above, it is necessary to evaluate coordination relations in the data. Recent studies have demonstrated that language-specific coordination relations between articulatory gestures can be reliably identified in the speech signal because of how they structure temporal variability (e.g. Gafos et al. 2014; Shaw 2022; Shaw et al. 2011). We illustrate this strategy with a simple model of gestural coordination for CC and CVC sequences. The framework for specifying the model builds on the model of articulatory representations proposed and deployed by various work (Gafos, 2002; Gafos et al., 2020; Shaw & Gafos, 2015), shown in Figure 1. We assume that a small number of gestural landmarks are available for coordination. In this case, the relevant landmarks are the gesture start, target, release, and end. The gesture start is the onset of movement associated with the gesture. The gesture target is the assumed goal of the movement. The gesture release is the onset of movement away from the assumed goal. The gesture end is the offset of controlled movement.

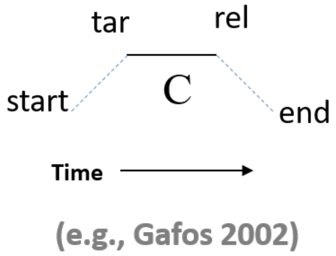
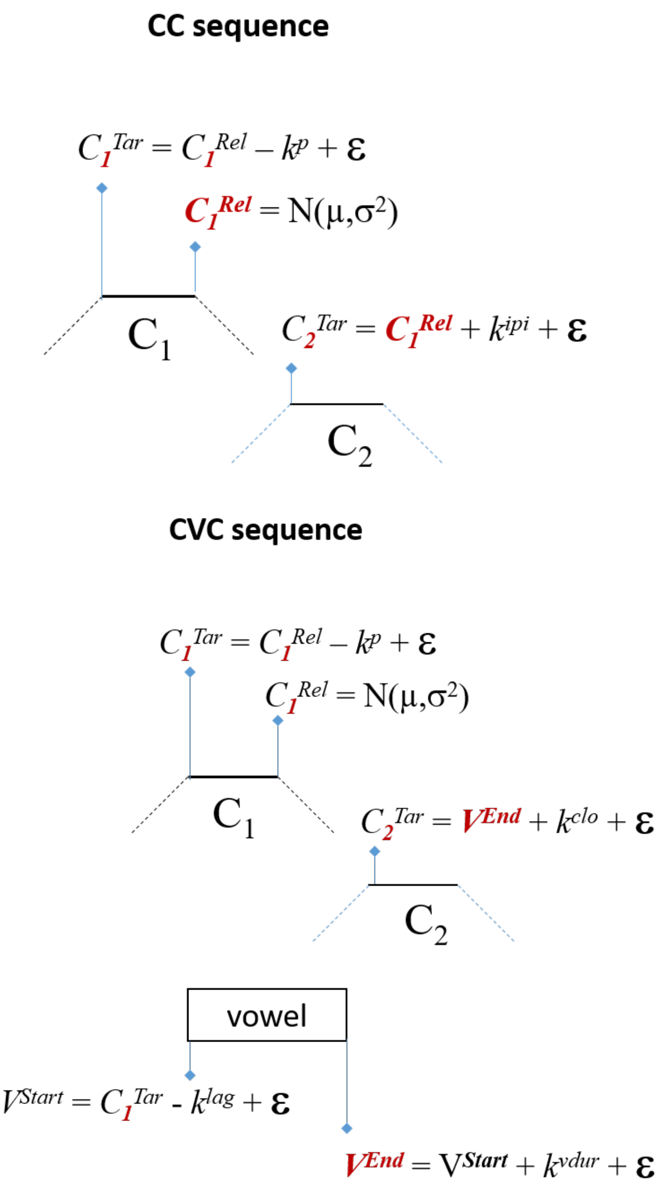


Figure 1: Four gestural landmarks posited by Gafos (2002) and subsequent work: the “start” of the gesture, sometimes also referred to as the “onset”, the (achievement of) “target”, abbreviated “tar”, the release (from constriction), abbreviated “rel”, and the “end” (of controlled movement), sometimes also referred to as “offset”.

In specifying stochastic models of gestural coordination, we define both inter- and intra-gestural timing as relationships between gestural landmarks (Shaw, 2022). The temporal precedence of intra-gestural landmarks is fixed: start → target → release → end. However, because we assume that gestures can temporally overlap, the inter-gestural relationship is not fixed. Rather, it depends on specification of inter-gestural coordination relations, which are often language specific. For example, consider a CV sequence. Within the set of gestural landmarks defined above, we could specify that the start of the vowel is coordinated with the start of the consonant. Alternatively, we could specify that the start of the vowel is coordinated with the target of the consonant, the release, or the offset. Alternatively, in this framework we could also specify that it is the target of the vowel (as opposed to the start) that is coordinated with the preceding consonant (see, e.g., Gafos et al. 2020; Roon et al. 2021; Shaw & Chen 2019). In some cases, coordination relations are known to map isomorphically to aspects of phonological structure, such as syllabic organization, making it possible to deduce higher level phonological structure from patterns of phonetic variability (e.g. Durvasula et al. 2021; Goldstein et al. 2007; Hermes et al. 2013, 2017; Shaw et al. 2009). Our focus here is on the relationship between coordination relations and kinematics. We build on the recent observation that different coordination relations (made available by the assumptions above) structure phonetic variability in different ways.



CC sequence

CVC sequence

Figure 2: Two coordination patterns, one for CC sequences and one for CVC sequences. A crucial difference involves the specification of inter-gestural coordination. For the CC sequence, the target of C_2 is timed to the release of C_1 (shown in red); for CVC, the target of C_2 is timed to the end of the vowel (shown in red); see text for complete description.

To illustrate this observation, we consider two different patterns of coordination, one for CC sequences and one for CVC sequences. An algorithm for generating gestural landmarks for each type of sequence is shown in Figure 2. The top shows a CC sequence in which the target of C_2 is timed directly to the release of C_1, shown in red. The phonetic constant, k^{ipi}, which could be zero, dictates how long after the release of C_1 the target of C_2 will occur, on average. The bottom panel shows a CVC sequence in which the target of C_2 is timed to the offset of the vowel (c.f., the release of C_1), shown in red. The other aspects of the coordination patterns are the same. In both examples, the target of C$_1$ is generated from a distribution defined by a constant, k^p (the p stands for plateau duration) and normally distributed error, and the release landmark, *rel*, is sampled from a normal distribution, N, defined by mean, μ, and variance, σ^2.

The result of the simulations are presented in Figure 3, which shows simulation results under very low levels of random noise, and Figure 4, which shows simulation results under levels of random noise typical of kinematic data.

Of interest is how variation in k^p differentially influences the interval between the release of C$_1$ to the target of C$_2$ (henceforth, ICI, for inter-consonantal interval). As k^p increases in CVC, ICI decreases. In contrast, for CC sequences in the bottom panel of Figure 2 , variation in k^p has no effect on ICI. Thus, a negative correlation between C$_1$ plateau duration and ICI is only consistent with the topology for the CVC sequence. This is regardless of the level of noise in the data. Figure 3 shows the same trend as Figure 4. The relation between C$_1$ duration and ICI is conditioned by the coordination relations between gestures, regardless of the degree of random variation added to the model. Since these two patterns of coordination make different predictions (Figures 3 and 4)—i.e., they structure variability in different ways—they can be diagnosed in the data.

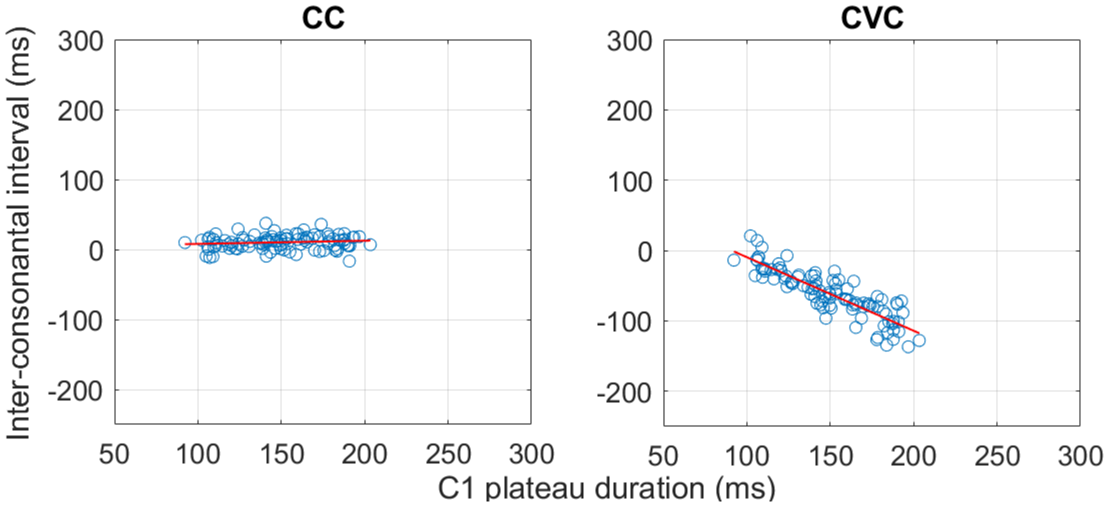


Figure 3: The simulated correlations between ICI and C$_1$ duration for the two coordination patterns in Figure 2 at low noise levels.

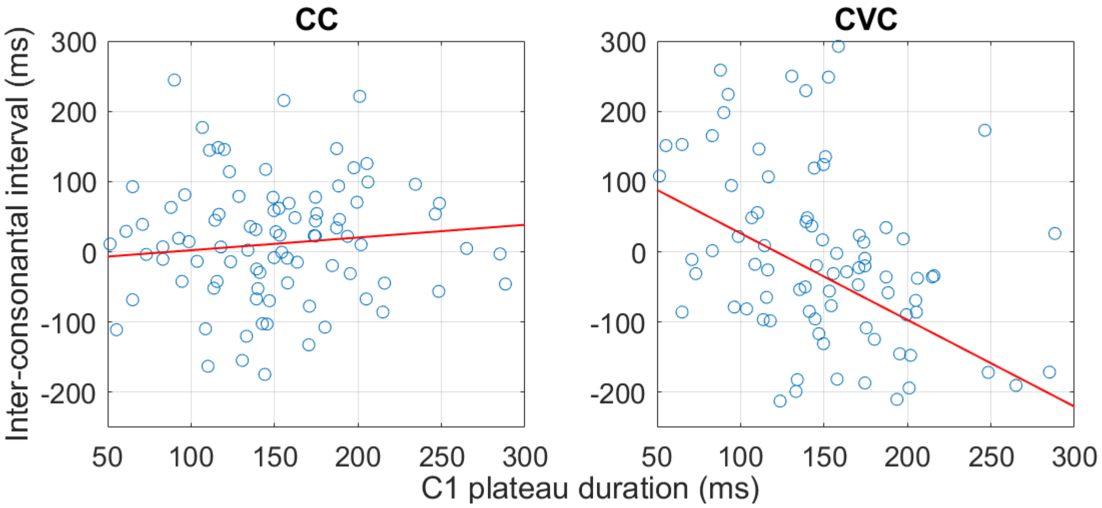


Figure 4: The simulated correlations between ICI and C_1 duration for the two coordination patterns in Figure 2 at noise levels typical of kinematic data.

The different coordination topologies in Figure 2 make different predictions about the covariation between C_1 duration and the inter-consonantal interval (ICI), defined as the interval from the release of C_1 to the achievement of target of C_2 (Shaw & Kawahara, 2018b). When the vowel is present (Figure 2, top), increases in C_1 duration will, all else equal, decrease ICI, because ICI is fixed in this coordination pattern. Thus, there should be a negative correlation between these intervals (C_1 duration and ICI) when the vowel is present. When the vowel is absent (Figure 2, bottom), on the other hand, variation in C_1 duration is not predicted to impact ICI, because the onset of C_2 is coordinated with the offset of C_1, i.e. C_2 onset can covary with C_1 offset. The rich theoretical landscape described above (Section 1.3) notwithstanding, these predictions follow what we take to be the standard view that gestural coordination is local and gesture duration triggers gestural reorganization.

Shaw & Kawahara (2018b) demonstrate that the different covariation patterns illustrated in 2 indeed hold in their dataset, implying that consonant clusters resulting from high vowel deletion are coordinated with each other. However, the dataset that was analyzed by Shaw & Kawahara (2018b) was somewhat limited, as the consonantal environments surrounding the devoiced/deleted vowels, which can crucially affect gestural reorganization, were not controlled in that experiment. Given that a larger and more controlled data set is available (Shaw & Kawahara, 2021), we aim to reexamine this question of how consonant clusters are organized after the intervening vowel is deleted.

2 Experimental methods

The data reported in this paper are based on those reported in Shaw & Kawahara (2021). Shaw & Kawahara (2021) established the probability of vowel deletion based on Bayesian classification of tongue dorsum trajectories. The aim of the current study is to assess the consequences of vowel deletion for the coordination of remaining gestures.

2.1 Participants

Seven adult native speakers of Tokyo Japanese participated in the experiment. All speakers were born in Tokyo, lived there at the time of their participation in the study, and had spent the majority of their lives there. Four speakers self-identified as male and three speakers self-identified as female. Participants were unaware of the purpose of the experiment and were compensated for their time and local travel expenses. Data from one speaker had to be excluded, because we were unable to record as many repetitions as other speakers. This speaker was originally coded as Speaker 6; their data is not discussed further below.

2.2 Stimuli

We analyze the same stimulus items which Shaw & Kawahara (2021) were able to classify in terms of vowel presence/absence. These items consist of two conditions based on the surrounding consonant types: fricative-stop (FS) and fricative-fricative (FF). The items are organized in dyads that differ in the status of the vowel, either voiced or devoiced (and possibly deleted). The 12 dyads are shown in Table 1.[4] All dyads consisted of two existing words in Japanese in which one member contained a C_1VC_2 sequence where both consonants are voiceless and the other member contained a minimally different C_1VC_2 sequence in which C_2 is voiced, hence V is not expected to devoice.

Table 1: The list of stimuli analyzed by Shaw & Kawahara (2021). S=Stop; F=Fricative. See footnote 4 for glosses. The first item of every pair contains /u/ in a devoicing environment; the second item contains /u/ in a voicing environment.

FS	FF
/ɸuton/ vs. /ɸudou/	/ɸusoku/ vs. /ɸuzoku/
/ɸutan/ vs. /ɸudan/	/ɸusai/ vs. /ɸuzai/
/ɸuta/ vs. /ɸuda/	/ɸusagaru/ vs. /ɸuzakeru/
/ʃutaisei/ vs. /ʃudaika/	/ʃusai/ vs. /ʃuzai/
/ʃutou/ vs. /ʃudou/	/ʃusa/ vs. /ʃuzan/
/ʃutokou/ vs. /ʃudouken/	/ʃuso/ vs. /ʃuzou/

2.3 Procedure

Each participant produced 14-15 repetitions of the target words in the carrier phrase: “okkee X to itte” (Ok, say X), where X is a stimulus word. Participants were instructed to speak as if they were making a request of a friend, in order to ensure that the speakers did not speak too formally or too slowly, which may inhibit vowel devoicing in the first place. This resulted in a corpus of 2,058 tokens (14 or 15 repetitions × 24 words × 6 speakers).

---

[4] The glosses are as follows. FF: blanket vs. not moving, burden vs. usual, top vs. amulet, subjectivity vs. thematization, FOOD NAME vs. hand-moving, Tokyo Highway vs.initiative; FS: shortage vs. attachment, debt vs. absence, filled vs. joke, organize vs. data collection, chair vs. abacus, main complaint vs. *sake*-making.

2.4 Equipment

We used an NDI Wave ElectroMagnetic Articulograph system sampling at 100 Hz to capture articulatory movement. NDI wave 5DoF sensors (receiver coils) were attached to three locations on the sagittal midline of the tongue, and on the lips, jaw (below the lower incisor), nasion and left/right mastoids. The most anterior sensor on the tongue, henceforth TT, was attached less than one cm from the tongue tip (see Figure 5). The most posterior sensor, henceforth TD, was attached as far back as was comfortable for the participant. A third sensor, henceforth TB, was placed on the tongue body roughly equidistant between the TT and TD sensors. Sensors were attached with a combination of surgical glue and ketac dental adhesive. Acoustic data were recorded simultaneously at 22 KHz with a Schoeps MK 41S supercardioid microphone (with Schoeps CMC 6 Ug power module).

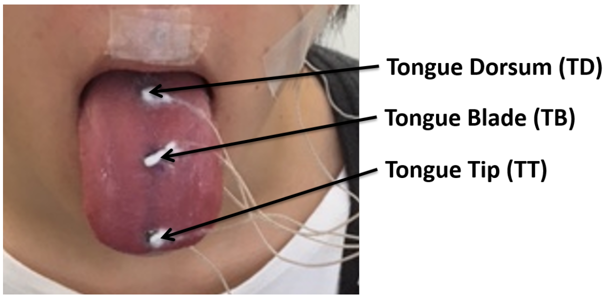


Figure 5: Illustration of the sensor placement (reproduced with permission from Shaw & Kawahara 2018b).

2.5 Stimulus display

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

2.6 Post-processing

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

3 Data analysis

3.1 Data processing

The wav files recorded in the experiment were submitted to forced alignment, using FAVE.[5] Textgrids from forced alignment were hand-corrected and, during this process, the target vowels were coded for voicing. Many vowels in devoicing environments were in fact devoiced, as evident from visual inspection of the spectrogram and waveform (see Shaw & Kawahara 2021). However, some tokens in the devoicing environment retained clear signs of glottal vibration. These vowels were coded as voiced, and excluded from further analysis. There were a total 240 vowels (12% of the data) in voiceless environments produced with some voicing; most of these 184/240 (77%) came from the FF condition but there were also 56/240 (23%) in the FS condition.

Articulatory data corresponding to each token were extracted based on the textgrids. To eliminate high frequency noise in the EMA recording, the kinematic data were smoothed using the robust smoothing algorithm (Garcia, 2010) and, subsequently, visualized in MVIEW, a Matlab-based program to analyze articulatory data (Tiede, 2005). Within MVIEW, gestural landmarks were parsed using the `findgest` algorithm. `Findgest` identifies gesture landmarks semi-automatically based upon the velocity signal in the movement toward and away from constrictions. The algorithm is semi-automatic in that it requires the user to identify the constriction of interest in one of the articulator movement trajectories. We identified gesture constrictions based on the primary oral articulator for each consonant: for the alveolar stops, /t/ and /d/, we used the tongue tip sensor; for the bilabial fricative, we used the lower lip sensor; for the alveolo-palatal fricative, we used the tongue blade sensor.

Whether to compute velocity signals based on movement in a single dimension, i.e., a component velocity, such as the vertical movement of the lower lip, or to instead refer to tangential velocity, a velocity signal that incorporates movement in all three available dimensions: vertical (up ↔ down), longitudinal (front ↔ back) and lateral (left ↔ right) is a researcher degree of freedom. Within the literature on kinematic analysis of speech movements, both approaches are common. Tangential velocity is preferable when the achievement of a speech production goal is distributed across dimensions: for example, a tongue tip movement to the alveolar ridge may involve both raising (vertical dimension) and also fronting (longitudinal dimension) of the tongue tip. If movements in both dimensions are in the service of achieving a single gestural goal, parsing the gesture based on just one dimension of movement may under-estimate movement velocity, which can impact gestural landmarks based on velocity-referential heuristics. On the other hand, there are cases in which a controlled movement can be better isolated by picking out a single movement dimension. Consider the case in which vertical movement is driven by one gesture while movement in the longitudinal dimension is driven by a temporally overlapped but distinct gesture. In this case, landmarks for each gesture would be better estimated by component velocities than by tangential velocities.

In our data it was generally appropriate to use tangential velocities, incorporating movement in three-dimensions into the gesture parse (below we discuss exceptions to this trend). Generally, there was very little movement in the lateral (left ↔ right) dimension, so tangential velocities were dictated primarily

---

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

by movement in the vertical and longitudinal dimensions. An example of a gesture parse of a bilabial fricative based on tangential velocity is provided in Figure 6. The top three panels show movement of the lower lip (LL) in the: (from top to bottom) longitudinal, vertical, and lateral dimensions. The bottom panel (red trajectory) shows the tangential velocity. The greatest displacement of the lower lip is in the vertical dimension, a movement magnitude of around 8 mm. However, there is also movement in the longitudinal dimension, i.e., lip protrusion, of about 3 mm and a small displacement, about 1 mm, in the lateral dimension. The bottom panel shows a sequence of four gestural landmarks, identified with reference to the tangential velocity signal, following, e.g., Shaw et al. (2009, 2011), Shaw et al. (2021), Shaw (2022): the "start" of the gesture, the achievement of "target", the "release" from constriction and the "end" of the gesture. Following past work, these landmarks were labeled with reference to the tangential velocity signal. The "start" landmark is when the velocity of the movement towards the constriction reaches 20% of peak velocity. The "target" landmark is labeled when velocity again lowers from its peak value to 20% of its peak value. Thus, the "start" and "target" landmarks are found on each side of the velocity peak. The "release" and "end" landmarks are identified with reference to the velocity peak in the movement away from constriction. The "release" landmark is labeled before the velocity peak, when velocity reaches 20% of its peak value. Finally, the "end' of the gesture is identified after the velocity peak in the movement away from constriction, at the time when velocity falls below 20% of its peak value. Gesture landmarks identified with reference to thresholds of peak velocity, as opposed to, e.g., velocity extrema (maximum and minimum), have the advantage of being generally more robust to small variations in spatial position than to velocity minima and maxima (see, e.g., Blackwood Ximenes et al. 2017 for discussion).

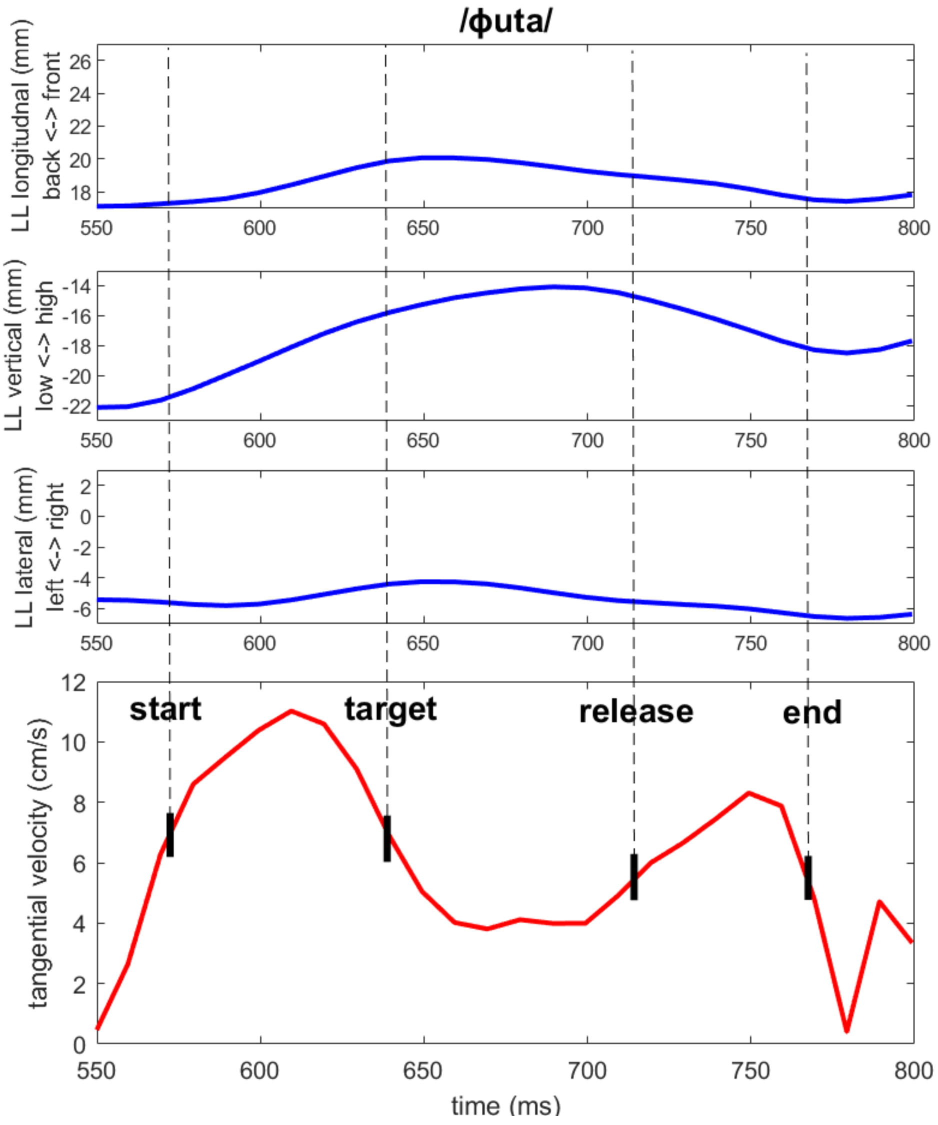


Figure 6: A sample articulatory trajectory and how the articulatory landmarks were identified using `findgest`.

Since we labelled tokens one at a time, we could observe when the application of the `Findgest` algorithm yielded an unrealistic gesture parse. There were two main reasons for this. Some tokens had velocity peaks that were not large enough to clearly parse out movement related to the consonants. If the local velocity peaks for either consonant were too small to detect gestural landmarks, we excluded the token from further analysis. A total of 239 tokens (13% of the data), 142 (7.8%) from the FS condition and 97 (5.3%) from the FF condition, were excluded for this reason. The resulting data set consisted of 1,579 tokens for analysis, which had clearly distinguishable consonantal gestures flanking the target vowel. Additionally, in some cases it was clear that the tangential velocity was inappropriately summing over multiple gestures. This was typically because a movement associated with C$_2$ overlapped with C$_1$. In these cases, we reverted to using component velocities instead of tangential velocities so as to disentangle the influence of overlapping gestures on the kinematics. For example, movement towards C_2 in one dimension, such as anterior movement of the tongue for /t/ in /ʃutaiseː/ sometimes overlapped in time with movement in another dimension associated with C_1, such as lowering of the tongue for /ʃ/. For this kind of case, we were able to isolate distinct velocity peaks for C_1 and C_2 by focusing on the primary spatial dimension of movement for each gesture: e.g., tongue lowering for /ʃ/ and tongue fronting toward the target for /t/. This approach is suggested in Guidelines for using MVIEW (Gafos et al., 2010b) and allowed us to consider a greater number of tokens for analysis. Instead of excluding tokens for which tangential velocities inappropriately summed movement components across distinct gestures, we instead parsed gestural landmarks in these cases using component velocities. For labial C_1, we used tangential velocity for 747 out of 783 tokens (95%); for coronal C_1, we used tangential velocity for 517 out of 796 tokens (65%).

The gestural landmarks parsed from the signal were used to define key measurements for further analysis. The inter-consonantal interval (ICI) was defined as the interval from the release of C_1 to the target of C_2 (see also §1.4). We defined C_1 plateau duration as the interval from target to release. These intervals allow us to test the key prediction laid out in §1.4 that the presence of a vowel conditions a negative correlation between them. Before conducting any analysis we removed outliers more than 2.5 standard deviations from the mean for these two key variables, C_1 plateau duration and ICI.

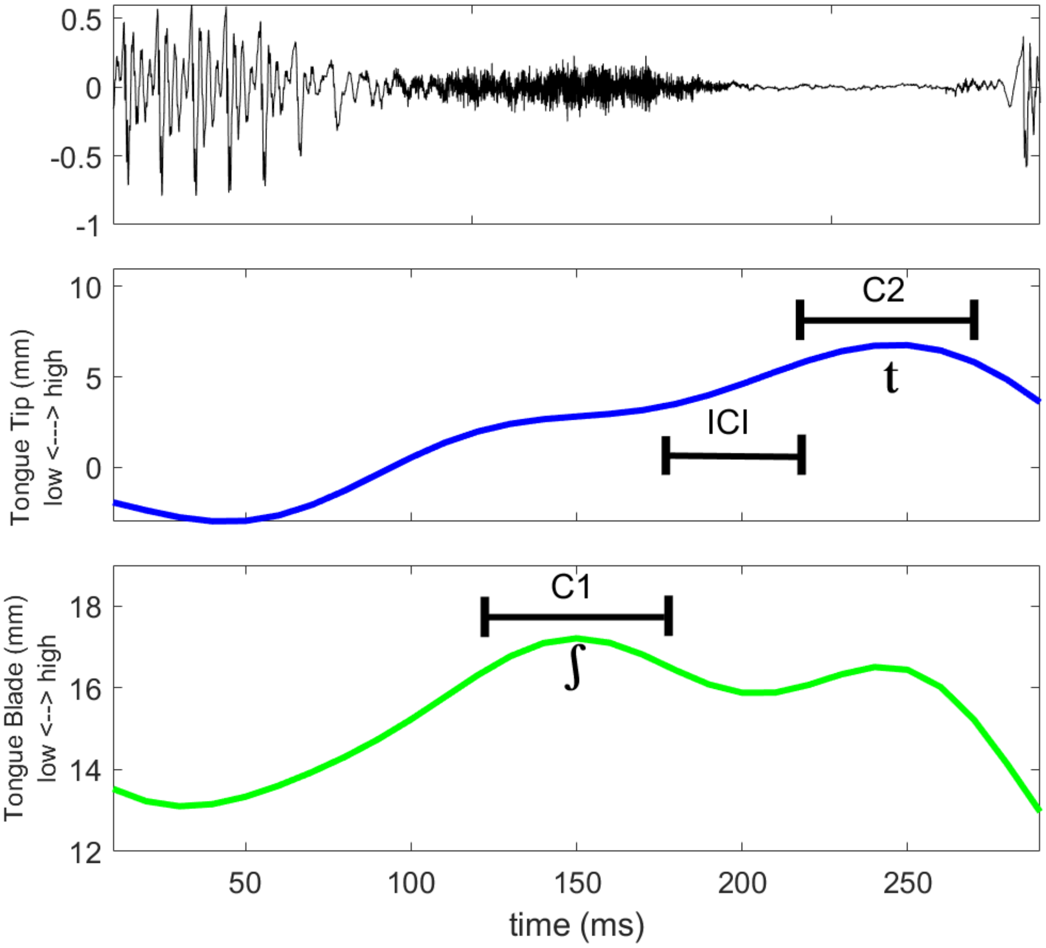


Figure 7: Illustrations of critical intervals. C_1 duration and C_2 duration are defined as the interval from target to release. ICI (Inter-consonantal Interval) is from the release of C_1 to the target of C_2

3.2 Assessing the probability of vowel deletion

The data that we are working with has already been classified for vowel presence/absense on the basis of the tongue dorsum trajectory, results reported in Shaw & Kawahara (2021) (for method, see also Shaw & Kawahara 2018a). For completeness, we briefly summarize the method here.

The temporal interval spanning from the start of movement of C_1, the consonant preceding the target vowel, and the end of movement of C_2, the consonant following the target vowel, was used to determine the probability of vowel deletion. We applied Discrete Cosine Transform (DCT) to represent the kinematic signal as the sum of four DCT components. Gaussian distributions over the DCT components for voiced vowel tokens were used to define a stochastic generator of vowel-present trajectories. We also setup a stochastic generator for the vowel-absent case. For each token of a devoiced item, we fit DCT components to the straight line connecting the position of the tongue dorsum at the onset and offset of the analysis window. The average of these DCT components (fit to the linear interpolation) defines the mean of the probability distribution for the "target absent" hypothesis. The standard deviation of the distributions is computed from the devoiced trajectories in the same manner as for the voiced items. Consequently, the probability distributions that characterize the “target absent” hypothesis are defined by linear interpolation and the variability around each DCT component in the data. We then used these two stochastically defined hypotheses—for target present and target absent trajectories—to classify the trajectories of devoiced items.

As the final step of the computational analysis, for each devoiced token, we determined the posterior probability of a vowel target, based on Bayesian classification of the tongue dorsum trajectory. The classifier was trained on the distributions described above for voiced tokens, which unambiguously contain a vowel target, and a noisy null hypothesis, defined as linear interpolation across the target interval. We do not force a categorical decision, but instead interpret the posterior probability of target absence for each token.

4 Results

Our main analysis compares tokens that have already been classified as containing a vowel or not. The classification results are reported in Shaw & Kawahara (2021). Here, we focus on the coordination of the consonants in tokens with and without a vowel.

We begin by reporting the inter-consonantal interval (ICI). Figure 8 shows the ICI by initial consonant (C1) place of articulation (PoA), coronal [ʃ] on the left and labial [ɸ] (“f”) on the right, and also by C_1C_2 manner sequence (ManSeq): fricative-fricative (FF) vs. fricative-stop (FS). Since the figure collapses across speakers, we present a z-score-normalized ICI here (see below for millisecond values by speaker). For the labial [ɸ]-initial clusters, there is little effect of manner sequence on ICI. For coronal [ʃ]-initial clusters, there is a trend towards longer ICI for FF than for FS clusters. However, the distributions are also less smooth for coronals [ʃ] than for labials [ɸ], which may indicate greater individual differences by speaker and/or by item for the tokens that begin with coronal fricatives.

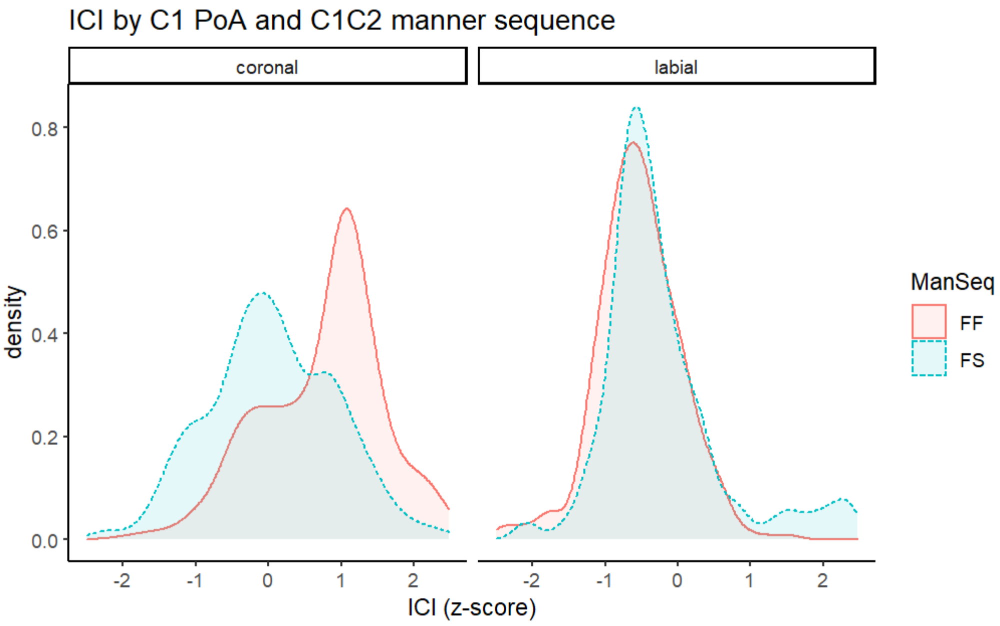


Figure 8: The distribution of inter-consonantal interval (ICI) values by C_1 place of articulation (PoA) and manner sequence (ManSeq). FF = Fricative-Fricative; FS = Fricative-Stop.

Figure 9 shows ICI in milliseconds (ms) by speaker, comparing voiced and devoiced environments. The voiced environments are those with voiced C_2 while the devoiced environments are those with voiceless C_2 (see Table 1). Although the distributions of ICI are generally not smooth, indicating variation across tokens (and items), there is heavy overlap between voiced and devoiced tokens. This indicates some degree of temporal preservation of ICI under devoicing. From the perspective of ICI, it seems that vowel devoicing does not entail vowel deletion. There were many devoiced tokens classified as containing a full vowel, just like voiced tokens. To assess the effect of vowel deletion, we need to incorporate the results on tongue dorsum movement classification.

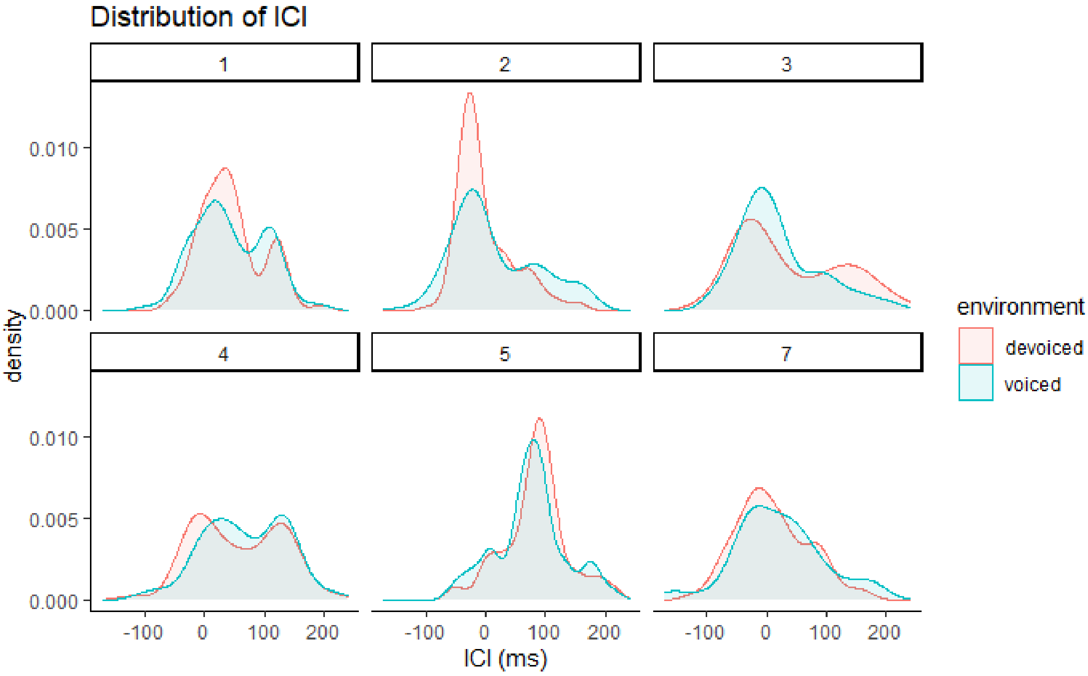


Figure 9: Distribution of ICI for each speaker.

Since the aim of this paper is to assess the effect of vowel deletion on the coordination of the remaining gestures, we took a conservative approach to interpreting the posterior probabilities reported in Shaw Kawahara (2021). We coded tokens with a greater than 0.95 probability of vowel deletion as "vowel absent", CC, and tokens with less than a 0.05 probability of vowel deletion as "vowel present", CVC. This reduces the amount of the data by 25%—from 526 tokens to 396 tokens. That is, 75% of the data is at the extreme ends of the probability distribution, indicating either a very low probability of deletion or a very high probability of deletion.

The main result is shown in Figure 10. This figure shows a scatter plot crossing two main conditions: manner sequence (FF vs. FS) and vowel presence (CC vs. CVC). Each panel plots the inter-consonantal interval (ICI) by C_1 duration. Recall the prediction from Figures 3 and 4 in §1.4. When a vowel is present we expect a negative correlation; increases in C_1 duration condition shorter ICI. We observe this negative correlation in three out of the four panels (all but the upper right panel). This is expected for CVC (bottom panels). We formulated three hypotheses about what would happen in CC (top panels). The results show that the negative correlation is observed in the FF items but not in the FS items.

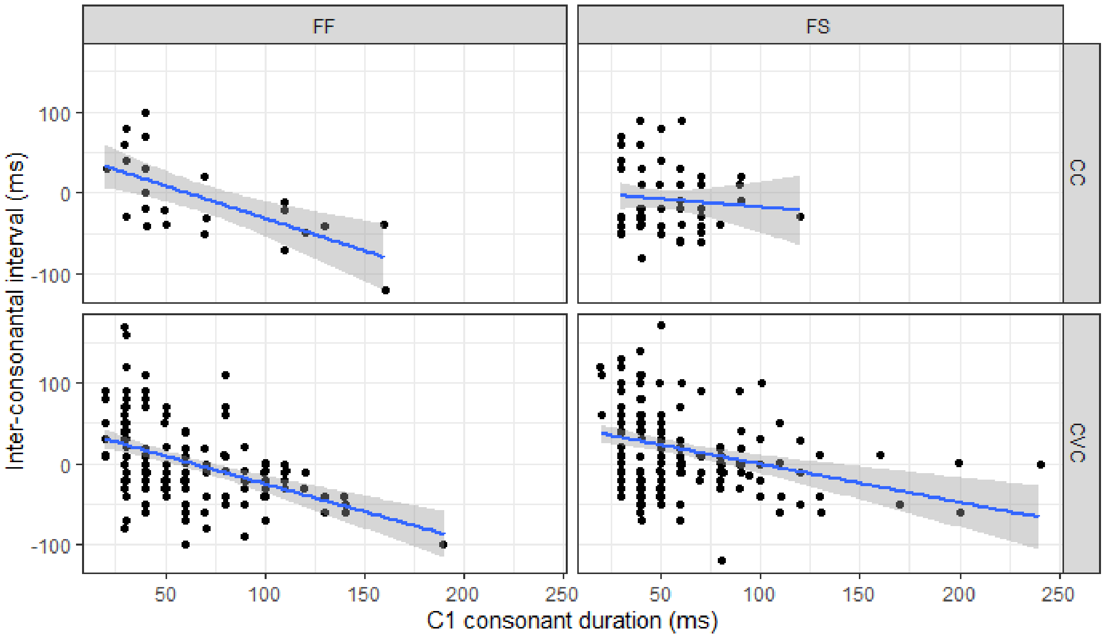


Figure 10: The observed correlations between ICIs and C_1 duration. Left=fricative-fricative condition; right=fricative-stop condition. Top=targetless tokens; bottom=CVC tokens.

To statistically assess the difference between FF and FS, we fit Bayesian regression models to z-scored ICI using the `brms` package (Bürkner, 2017) in R (version 4.1.3). Since we seek to evaluate statistically whether the effect of C_1 duration on ICI is modulated by manner sequence (FF vs. FS consonant clusters), we are interested only in the CC tokens. We therefore fit a model to just the data in the upper panels of Figure 10.

The model contained a random intercept for speaker and a random by-speaker slope for manner sequence (FF vs. FS). The fixed factors were C1 place of articulation (PoA), either labial or coronal, z-score normalized C_1 duration, and manner sequence (FF vs. FS), along with the two-way interactions between C_1 identity and manner sequence (ManSeq) and between C_1 duration and manner sequence (ManSeq). The formula for the model is given in (1) below.

(1) $zICI \sim zC1\_duration * ManSeq + PoA * ManSeq + (1|speaker) + (0 + ManSeq|speaker)$

The procedure for fitting the models followed recommendations of learnB4SS (version S 1.0.7.9000), the LabPhon-sponsored workshop on Bayesian regression for Speech Sciences[6]. All priors were set to be weakly informative (Gelman et al., 2018): the priors for fixed factors drew from a normal distribution with a mean of 0 and standard deviation of 2; the random effects drew from a cauchy distribution with a mean of 0 and standard deviation of 0.1. We ran four chains with 2,000 warmups and an additional 1,000 samples. There were no divergent transitions. Additionally, the $\hat{R}$-values, a diagnostic for conver-

---

[6] `https://learnb4ss.github.io/`

gence, for all fixed effects were 1.0, indicating that chains mixed successfully. See the markdown file for complete details, which is available at `osf.io/gmr8j`.

Figure 11 provides a graphical representation of the model results, showing ranges of values that each estimated parameter can take. For each fixed factor, the plot shows the uncertainty around the model estimates. The 95% credible interval (CrI) is shown as a shaded interval; the tails beyond 95% credible intervals are unshaded.

All of the probability mass for consonant plateau duration $zC1\_duration$ is negative ($\beta = -0.50$, 95% CrI=[-0.76, -0.23]), indicating a highly reliable effect. As C1 duration increases, ICI decreases. The effect of C1 place of articulation PoA, i.e., labial [ɸ] vs. coronal [ʃ], is negative, indicating that ICI is shorter following labials than following coronals, but the thick portion of the distribution overlaps with zero ($\beta = -0.28$, 95% CrI=[-0.81, 0.24]). This indicates that PoA does not have a reliable effect on ICI. The same goes for the manner sequence factor, $ManSeq$. ICI is somewhat shorter following FS than FF, but this effect of $ManSeq$ is not very credible ($\beta = -0.18$, 95% CrI=[-0.68, 0.35]). The interaction between $ManSeq$ and PoA tends to be positive but also overlaps zero substantially ($\beta = 0.16$, 95% CrI=[-0.43, 0.70]). Finally, we turn to the interaction between $zC_1\_duration$ and $ManSeq$, the factor most relevant to our theoretical hypotheses. The entire thick portion of this distribution was positive, suggesting that this factor is meaningful ($\beta = 0.43$, 95% CrI=[0.12, 0.72]). The direction of this effect functions to cancel out the main effect of consonant duration in the FS environment. That is, the FS items are a reliable exception to the general trend: a negative influence of C1 duration on ICI.

In short, the effect of C_1 duration on ICI is modulated by manner sequence (FF vs. FS), as indicated by the meaningful interaction term. The negative effect of consonant duration predicted for CVC (§1.3) and verified in our data (Figure 10, bottom) persists even in CC, but only when both consonants are fricatives. In FS sequences, vowel deletion seems to have resulted in gesture reorganization.

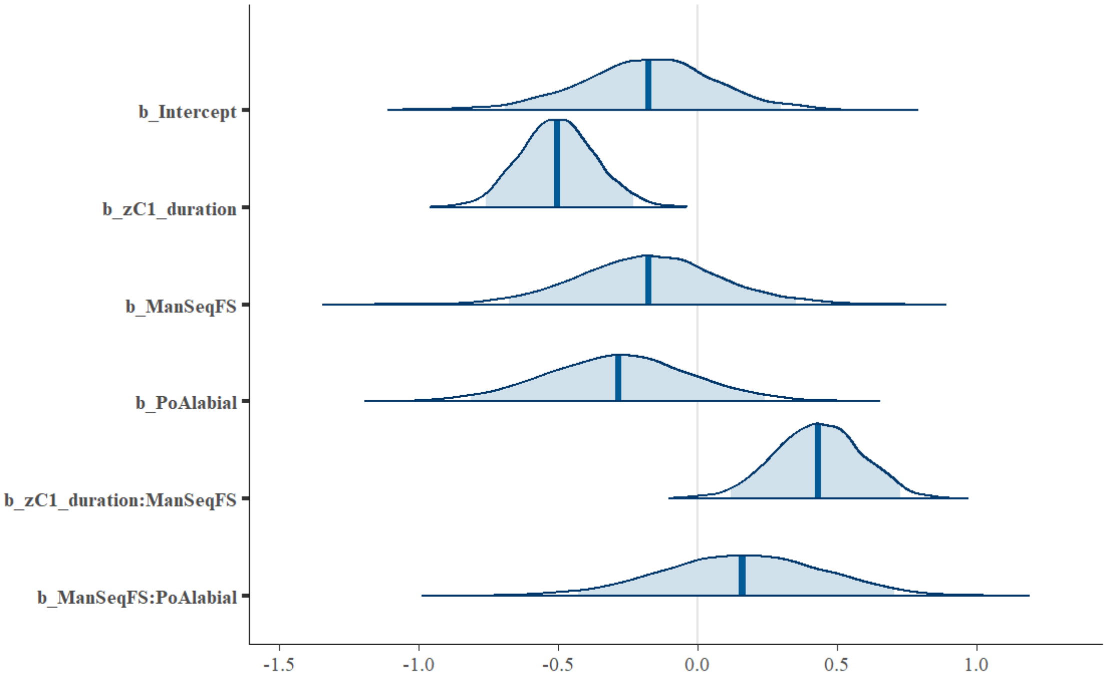


Figure 11: Posterior probability distributions of each estimated parameter. The shaded portion of the distribution covers 95% of the estimates.

The statistical results confirm the pattern in the top two panels of Figure 10. There is a negative effect of C1 duration on ICI for FF sequences (Figure 10: left) but not for FS sequences (Figure 10: right).

We next evaluate whether the effect of C1 on ICI found for FF sequences is the same for items with (CVC) and without (CC) a vowel. To do this we fit a Bayesian regression model to the FF data. As above, we included fixed effects of C1 duration $zC1\_duration$ and place of articulation PoA and a random intercept for speaker. We also included a fixed effect of vowel presence/absence $vowel$, so that we effectively compare the top and bottom left panels of Figure 10 along with a by-speaker random slope for $vowel$. The formula for the model is given below:

(2) $$zICI \sim zC1\_duration * vowel + PoA * vowel + (1|speaker) + (0 + vowel|speaker)$$

As expected from the figure, the main effect of $zC1\_duration$ was negative and did not overlap with zero ($\beta = -0.49$, 95% CrI=[-0.78, -0.21]). The interaction between $zC1\_duration$ and $vowel$ was weakly positive and heavily overlapped with zero ($\beta = 0.14$, 95% CrI=[-0.16, 0.44]). This indicates that the pattern for fricative-fricative and fricative-vowel-fricative items is not appreciably different. For both types of items there is a strong negative effect of C1 on ICI.

For completeness, we also evaluate the effect of vowel presence/absence on FS sequences, again using the formula in (2) above. In this case, the main effect of $zC1\_duration$ trended negative but was weaker ($\beta = -0.12$, c.f., -0.49 above) and not reliable, with the credible interval overlapping zero substantially: ($\beta = -0.12$, 95% CrI=[-0.33, 0.11]). The interaction between $zC1\_duration$ and $vowel$, however, was much stronger ($\beta = -0.21$) and more credible with only small amount of probability mass overlapping zero: ($\beta = -0.21$, 95% CrI=[-0.44, 0.02]). The negative effect of C1 duration on ICI is much stronger in items in which a vowel was identified (CVC) than in items in which a vowel is absent.

The statistical analyses above confirm the trends in Figure 10. As predicted by our simulations (Figures 3 and 4), C_1 duration has a negative effect on ICI when there is a vowel intervening between consonants, i.e., CVC items. For CC items, those classified as lacking a vowel, fricative-fricative sequences differed from fricative-stop sequences. Only fricative-stop sequences showed the pattern predicted for CC (Figures 3 and 4). In contrast, fricative-fricative sequences showed a timing pattern indistinguishable from CVC, despite lacking a tongue dorsum movement for the vowel.

5 Discussion

We investigated whether gesture reorganization accompanies vowel deletion, making use of a data set for which tokens have already been classified as containing a vowel or not (Shaw & Kawahara, 2021). Based on past research, we motivated three competing hypotheses (Section 1.3): (1) that vowel deletion triggers reorganization of gestural coordination; (2) that gesture coordination is unaffected by vowel deletion; and (3) that gestural reorganization depends on consonant context. Our results supported the third hypothesis: we found gestural reorganization in fricative-stop (FS) clusters but not fricative-fricative (FF) clusters. Past work established that these two environments show vowel deletion with similar frequency (Shaw & Kawahara, 2021), which was established by classifying tongue dorsum trajectories. Even though there is not a significant difference in deletion probability in these two environments, the current study showed that there is a difference in terms of gestural coordination.

Gestural reorganization is conditioned by consonant environment. Specifically, we observed gestural reorganization when vowel deletion results in fricative-stop clusters (FS). In contrast, fricative-fricative (FF) clusters resisted gestural reorganization, showing the same coordination pattern as CVC (vowel present) sequences. A key implication of our results is that temporal structure may be preserved even when there is no articulatory displacement, at least in some phonological environments.[7]

Importantly, the lack of a vowel target in these data was not due to target undershoot. Shaw & Kawahara (2021) examined this possibility in depth and ultimately rejected it. Many tokens classified as lacking a vowel were amongst the longest durations in the data. Thus, these speakers produced vowels without a target even when not under time pressure. That temporal structure may be preserved even without overt articulatory movement is consistent as well with patterns of compensatory lengthening, whereby the loss of a segment does not alter the temporal structure of a higher level constituent, e.g, word (Kavitskaya, 2002).

In motivating our hypotheses, we explored a number of theoretical possibilities for how coordination

---

[7] An anonymous reviewer points out a possible alternative line of explanation, locating the difference between FF and FS conditions in our data in the articulatory differences between fricatives and stops, including, possibly, differential contributions of the jaw in producing these consonants. Although beyond the scope of our study, which has a different theoretical motivation, We view this as an interesting line of inquiry for future exploration.

could be maintained following vowel deletion: (1) selection/coordination theory (Tilsen, 2016), (2) non-local timing mechanisms, such as a moraic or syllable-level clock (Barbosa, 2007; O'Dell & Nieminen, 2019) and (3) the gestural analog of a "ghost segment" or gradient symbolic representation (Hsu, 2019; Lionnet, 2017; Smolensky & Goldrick, 2016; Walker, 2020; Zimmermann, 2019).

Each of these theories can, in principle, deal with the maintenance of a gestural coordination pattern in the absence of a vowel, or, at least, in the absence of a vowel movement detectable in the kinematic signal. However, none of them are particularly well-suited to explaining the difference between fricative-fricative (FF) and fricative-stop (FS) environments. A key theoretical implication of our results is that any one of these accounts would require some augmentation. One possibility, which we pursue here in some detail, is that variable devoicing in fricative-fricative (FF) environments is related to the maintenance of gestural coordination.

In the introduction, we pointed out that high vowel devoicing is less likely in fricative-fricative environments than in stop-stop or fricative-stop environments (Fujimoto, 2015; Maekawa & Kikuchi, 2005). Putting this together with our data, the environment less likely to show devoicing (FF) is also the environment that resists gestural reorganization, maintaining the temporal structure of fricative-vowel-fricative strings even in the absence of other acoustic or kinematic reflexes of the vowel. It may be possible to link these two facts about FF environments. Given the variability of devoicing in the FF environment, Japanese listeners will experience voiced vowels in FF environments more than in FS environments. This experience of a voiced vowel could encourage a higher degree of vowel activation in F_F than in F_S contexts. That is, as compared to vowels that are deleted (i.e., vowels that have no activation), weak activation of vowels in fricative-vowel-fricative may be reinforced by occasionally hearing voiced vowels in this context.

In the interest of fleshing out this idea, we consider how cases of timing preservation could be understood in terms of a weakly activated gesture, i.e., a gesture that is present but activated weakly enough that its kinematic reflexes cannot be observed. There may additionally be a connection between weakly active *gestures* and weakly activated *segments*, as in gradient symbolic representations (Smolensky & Goldrick, 2016) and conceptually-related proposals (Hsu, 2019; Lionnet, 2017; Walker, 2020; Zimmermann, 2019). Although the details of the proposals vary, a common theme is that degrees of activation of phonological representations have consequences for phonological computation. In some cases, evidence for the weakly active segment may surface only in its impact on phonological computation. By analogy, evidence for a weakly active gesture may exist only in its impact on the timing of other gestures.

Whether segments found to be gradiently active for the purposes of phonological computation also impact temporal organization remains an open question. For example, do liaison consonants—argued to be gradiently active (Smolensky & Goldrick, 2016)—also block gestural reorganization? There are already some proposals linking gradient activation of segments to degrees of gestural activation. For example, reduced activation at the segment level has been argued to impact gesture activation duration, in models of speech errors (Goldrick & Chu 2014, c.f. Stern et al. 2022). Extreme reduction could make the gesture undetectable in the kinematic record and yet still present for the purpose of conditioning coordination relations between other gestures.

To derive our results, some new assumptions are required. The first is that gradient gesture activation is related to the probability of surface occurrence, based on perception. Additionally, we assume that a voiced vowel provides less uncertainty about surface occurrence of a lingual gesture than a devoiced vowel. Another assumption is that a partially active gesture can condition coordination patterns with other gestures. On these assumptions, the degree of activation of high back vowels in Tokyo Japanese may be systematically higher in fricative-fricative contexts than in fricative-stop contexts, by virtue of the occasional failure of high vowel devoicing in this context. When it comes to articulation, partial activation is sufficient for coordination with other gestures even when insufficient to drive the articulators towards a vowel-specific target.

Although we opted to outline this proposal in terms of gradient activation as opposed to other theories that could also be augmented to explain the results, there are other cases in which loss of a surface gesture preserves timing. Intervocalic velar stops in Iwaidja can be lenited completely. However, lenition of the stop in /aka/ yields a vocalic interval that is greater than two times the duration of stressed /a/, suggesting that some temporal aspect of the deleted consonant remains (Shaw et al., 2020). Another case comes from Tibetan (Geissler, 2021), in which syllables with lexical tones have been shown to have a pattern of C-V coordination that is distinct from C-V coordination in toneless syllables. Specifically, the vowel starts later in time relative to the consonant in syllables with lexical tone (Mandarin: Shaw & Chen 2019; Zhang et al. 2019; Thai: Karlin & Tilsen 2015; Swedish: Svensson Lundmark et al. 2021). Some speakers of Tibetan who do not produce a lexical tone contrast maintain the C-V coordination characteristic of tonal syllables.

To the extent that weak activation of a vowel in production maintains temporal structure, it may also facilitate comprehension. High vowel devoicing, although detrimental to phoneme spotting, actually facilitates lexical retrieval of real words relative to fully voiced vowels in devoicing environments (Ogasawara & Warner, 2009; Ogasawara, 2013). In the word spotting task, complete vowel deletion, tested by splicing out a vowel from the acoustic signal, hinders performance, even when the vowel is spliced from a devoicing context (Cutler et al., 2009). There appears to be a difference between devoicing and deletion in comprehension. Our study indicates that there is an intermediate possibility between vowel devoicing and full vowel deletion. Possibly, a weakly active vowel gesture in the FF environment resolves some tension between the application of a phonological process and the faithful production of a lexical item. The tension emerges from the perceptual experience of speakers, which may include some fricative-vowel-fricative sequences produced variably with a fully voiced vowel and with a devoiced vowel. Maintaining the temporal structure of a vowel through weak activation may also facilitate comprehension, although this speculation requires empirical testing.

The tension involved in FF sequences is reminiscent also of recent accounts of incomplete neutralization, in which maintaining consistent pronunciation of a word facilitates partial resistance to phonological processes (Braver, 2019; Yu, 2007). Japanese words with lexical pitch accent are sometimes produced with reduced or absent pitch contours. In wh-interrogative sentences, scope is signalled by the erradication of lexical pitch accents in words intervening between the wh-item and the complementizer (Deguchi & Kitagawa, 2002; Richards, 2010). However, we found that complete eradication is sometimes resisted, which may again reflect a tension between consistent production of a lexical item and a productive phonological process (Kawahara et al., 2022).

If the discussion above is on the right track, it suggests a connection between variable devoicing and a lack of gestural reorganization. More generally, weak gestural activation blocking reorganization might be more likely in environments in which the phonological process—in this case devoicing—is more variable. The assumption here is that more consistent devoicing, as observed in the FS context, provides less evidence for the presence of a vowel. On this account, the occasional absence of devoicing has the consequence of blocking gestural reorganization. The weakly activated gesture maintains the temporal structure of the vowel, without requiring spatial displacement, providing a compromise between competing pressures on articulation.

6 Conclusion

We investigated whether vowel deletion triggers reorganization of the remaining gestures, making use of variable vowel deletion in Tokyo Japanese. Our stimuli included vowels deleted in two consonant environments: fricative-fricative (FF) and fricative-stop (FS). Results indicated gestural reorganization only in the FS clusters and not in FF clusters. This indicates that deletion of a vowel does not necessarily result in gestural reorganization. The temporal structure of a word can be maintained even when a segment is lost. Possible theoretical mechanisms for maintaining timing in the face of deletion include weakly activated gestures and/or higher level clocks. The differences between FF and FS clusters may follow from the optionality of vowel devoicing—a prerequisite for deletion—in FF clusters.

Statement of Ethics

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

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions Statement

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

Acknowledgments

Thanks to Chika Takahashi and Jeff Moore for help with the experiment, to Emily Grabowski for coding the data for devoicing, to Noah Macey for parsing out gestures. This project is supported by JSPS grant #15F15715 to both authors.

References

Barbosa, Plínio. 2007. From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication* 49(9). 725–742.

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

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

Blackwood Ximenes, A, Jason A. Shaw & Christopher Carignan. 2017. A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English. *The Journal of the Acoustical Society of America* 142(1). 363–377.

Braver, Aaron. 2019. Modeling incomplete neutralisation with weighted phonetic constraints. *Phonology* 36(1). 1–36.

Browman, Catherine & Louis Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In Mary Beckman & John Kingston (eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech*, vol. 1, 341–397. Cambridge: Cambridge University Press.

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models using Stan. R package.

Chitoran, Ioana, Louis Goldstein & Dani Byrd. 2002. Gestural overlap and recoverability: Articulatory evidence from georgian. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology VII*, 419–447. Berlin: Mouton de Gruyter.

Chomsky, Noam & Moris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Crouch, Caroline, Argyro Katsika & Ioana Chitoran. 2020. The role of sonority profile and order of place of articulation on gestural overlap in Georgian. *Proceedings of Speech and Prosody 2020* .

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

Davidson, Lisa. 2006. Schwa elision in fast speech: Segmental deletion or gestural overlap. *Phonetica* 63(2-3). 79–112.

Deguchi, Masanori & Yoshihisa Kitagawa. 2002. Prosody and wh-questions. *Proceedings of NELS* 32. 73–92.

Durvasula, Karthik, Mohammed Qasem Ruthan, Sarah Heidenreich & Yen-Hwei Lin. 2021. Probing syllable structure through acoustic measurements: case studies on american english and jazani arabic. *Phonology* 38(2). 173 –202.

Ellis, Lucy & William Hardcastle. 2002. Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. *Journal of Phonetics* 30. 373–396.

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

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

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

Gafos, Adamantios. 1999. *The articulatory basis of locality in phonology*. New York: Garland.

Gafos, Adamantios. 2002. A grammar of gestural coordination. *Natural Language and Linguistic Theory* 20. 269–337.

Gafos, Adamantios, Simon Charlow, Jason Shaw & Phil Hoole. 2014. Stochastic time analysis of syllable-referential intervals and simplex onsets. *Journal of Phonetics* 44. 152–166.

Gafos, Adamantios, Philip Hoole, Kevin Roon & Chakir Zeroual. 2010a. Variation in timing and phonological grammar in moroccan arabic clusters. In Cécile Fougeron (ed.), *Laboratory phonology 10:*

Variation, detail and representation, Mouton De Gruyter.

Gafos, Adamantios, Christo Kirov & Jason Shaw. 2010b. Guidelines for using mview. Available at `https://www.ling.uni-potsdam.de/~gafos/courses/Rethymno/Guide_2.pdf`.

Gafos, Adamantios, Jens Roeser, Stavroula Sotiropoulou, Philip Hoole & Chakir Zeroual. 2020. Structure in mind, structure in vocal tract. *Natural Language and Linguistic Theory* 38. 43–75.

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

Geissler, Chris. 2021. *Temporal articulatory stability, phonological variation, and lexical contrast preservation in diaspora tibetan*: Yale University Doctoral dissertation.

Gelman, Andrew, A. Jakulin, M.G. Pittau & Y-S. Su. 2018. A weakly informative default prior distribution for logistic and other regression models. *Annual Applied Statistics* 1360–1383.

Goldrick, Matthew & K. Chu. 2014. Gradient co-activation and speech error articulation: Comment on pouplier and goldstein (2010). *Language, Cognition and Neuroscience* 29(4). 452–458.

Goldstein, Louis, Ioana Chitoran & Elisabeth Selkirk. 2007. Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber. Paper presented at the XVIth International Congress of Phonetic Sciences, Saabrucken, Germany.

Gouskova, Maria. 2004. Relational markedness in OT: The case of syllable contact. *Phonology* 21(2). 201–250.

Hermes, Anne, D. Mücke & B. Auris. 2017. The variability of syllable patterns in Tashlhiyt Berber and Polish. *Journal of Phonetics* 64. 127–144.

Hermes, Anne, D. Mücke & Martine Grice. 2013. Gestural coordination of Italian word-initial clusters: The case of "impure s". *Phonology* 30(1). 1–25.

Hsu, Brian. 2019. Exceptional prosodification effects revisited in Gradient Harmonic Grammar. *Phonology* 36. 225–263.

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

Kaisse, Ellen M. & Patricia Shaw. 1985. On the theory of lexical phonology. *Phonology* 2. 1–30.

Karlin, R. & Sam Tilsen. 2015. The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai. *POMA* 22. 060006.

Kavitskaya, Dasha. 2002. *Compensatory lengthening: Phonetics, phonology, diachrony*. New York: Routledge.

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

Kawahara, Shigeto, Jason A. Shaw & Shinichiro Ishihara. 2022. Assessing the prosodic licensing of wh-in-situ in japanese: A computational-experimental approach. *Natural Language and Linguistic Theory* .

Kochetov, Alexei & Marianne Pouplier. 2008. Phonetic variability and grammatical knowledge: an articulatory study of Korean place assimilation. *Phonology* 25(3). 399–431.

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

Lialiou, Maria, Stavroula Sotiropoulou & Adamantios Gafos. 2021. Spatiotemporal coordination in word-medial stop-lateral and s-stop clusters of American English. *Phonetica* .

Lionnet, Florian. 2017. A theory of subfeatural representations: The case of rounding harmony in Laal. *Phonology* 34(3). 523–564.

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

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

Nolan, Francis. 1992. The descriptive role of segments: Evidence from assimilation. In Gerard R. Docherty & Robert Ladd (eds.), *Papers in laboratory phonology II: Gesture, segment, prosody*, 261–280. Cambridge: Cambridge University Press.

O'Dell, Michael & T. Nieminen. 2019. Syllable rate, syllable complexity and speech tempo perception in Finnish. *Proceedings of the 19th International Congress of Phonetic Sciences* .

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

Ogasawara, Naomi & Natasha Warner. 2009. Processing missing vowels: Allophonic processing in japanese. *Language and Cognitive Processes* 24(3). 376–411.

Ohala, John & Haruko Kawasaki-Fukumori. 1997. Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In Stig Eliasson & Ernst Hakon Jahr (eds.), *Language and its ecology: essays in memory of einar haugen*, 343–365. Berlin: Mouton de Gruyter.

Pouplier, Marianne & Luis Goldstein. 2014. The relationship between planning and execution is more than duration: Response to Goldrick & Chu. *Language, Cognition and Neuroscience* 29(9). 1097 – 1099.

Richards, Norvin. 2010. *Uttering trees*. MIT Press.

Roon, K. D., P. Hoole, C. Zeroual, S. Du & A. I. Gafos. 2021. Stiffness and articulatory overlap in moroccan arabic consonant clusters. *Laboratory Phonology* .

Shaw, J. A., C. Carignan, T. G. Agostini, R. Mailhammer, M. Harvey & D. Derrick. 2020. Phonological contrast and phonetic variation: The case of velars in iwaidja. *Language* 96(3). 578–617.

Shaw, Jason. 2022. Microprosody. *Language and Linguistic Compass* .

Shaw, Jason & W.R. Chen. 2019. Spatially conditioned speech timing: Evidence and implications. *Frontiers in Psychology* 10. 2726.

Shaw, Jason, Karthik Durvasula & Alexei Kochetov. 2021. Articulatory coordination distinguishes complex segments from segment sequences. *Phonology* .

Shaw, Jason & Adamantios Gafos. 2015. Stochastic time models of syllable structure. *PLOS ONE* 10(5).

Shaw, Jason, Adamantios Gafos, Phil Hoole & Chakir Zeroual. 2009. Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology* 26(1). 187–215.

Shaw, Jason, Adamantios Gafos, Phil Hoole & Chakir Zeroual. 2011. Dynamic invariance in the phonetic expression of syllable structure: A case study of Moroccan Arabic consonant clusters. *Phonology* 28(3). 455–490.

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

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

Shaw, Jason & Shigeto Kawahara. 2021. More on the articulation of devoiced [u] in Tokyo Japanese: Effects of surrounding consonants. *Phonetica* 78(5/6). 467–513.

Smith, Caroline. 1995. Prosodic patterns in the coordination of vowel and consonant gestures. In Bruce Connell & Amalia Arvaniti (eds.), *Papers in laboratory phonology IV: Phonology and phonetic evidence*, 205–222. Cambridge: Cambridge University Press.

Smolensky, Paul & Matt Goldrick. 2016. Gradient symbolic representations in grammar: The case of French liaison. Ms. Johns Hopkins University and Northwestern University.

Sotiropoulou, S. & A Gafos. 2022. Phonetic indices of syllabic organization in german stop-lateral clusters. *Laboratory Phonology* .

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

Stern, M. C, M. Chaturvedi & J. A. Shaw. 2022. A dynamic neural field model of phonetic trace effects in speech errors. *Proceedings of the 44th Annual Conference of the Cognitive Science Society* .

Svensson Lundmark, Malin, Gilbert Ambrazaitis, Johan Frid & Susanne Schötz. 2021. Word-initial consonant—vowel coordination in a lexical pitch-accent language. *Phonetica* 78(5/6). 515–569.

Tiede, Mark. 2005. Mview. Software.

Tilsen, Sam. 2016. Selection and coordination: the articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55. 53–77.

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

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

Walker, Rachel. 2020. Gradient activity in Korean place assimilation. *Proceedings of NELS* 50. 207–220.

Yu, Alan. 2007. Understanding near mergers: The case of morphological tone in Cantonese. *Phonology* 24. 187–214.

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

Zimmermann, Eva. 2019. Gradient symbolic representations and the typology of ghost segments. *Proceedings of AMP 2018* .

Zsiga, Elizabeth. 2020. *The phonology/phonetics interface*. Edinburgh: Edinburgh University Press.