Recoverability-driven coarticulation:
Acoustic evidence from Japanese high vowel devoicing

Running title: *Recoverability and Japanese high vowel devoicing*

James Whang[a]
Department of Linguistics
New York University
10 Washington Place
New York, NY 10003

---

[a] email: james.whang@nyu.edu

Abstract

High vowel devoicing in Japanese, where /i, u/ in a C_1VC_2 sequence devoice when both C_1 and C_2 are voiceless, has been studied extensively, but factors that contribute to the devoiced vowels' likelihood of complete deletion is still debated. This study examines the effects of phonotactic predictability on the deletion of devoiced vowels. Native Tokyo Japanese speakers (N=22) were recorded in a sound-attenuated booth reading sentences containing lexical stimuli. C_1 of the stimuli were /k, ʃ/, after which either high vowel can occur, and /ʧ, ɸ, s, ç/, after which only one of the two occurs. C_2 was always a stop. C_1 duration and center of gravity (COG), the amplitude weighted mean of frequencies present in a signal, were measured. Duration results show that devoicing lengthens only non-fricatives, while it has either no effect or a shortening effect on fricatives. COG results show that coarticulatory effects of devoiced vowels are evident in /k, ʃ/ but not in /ʧ, ɸ, s, ç/. Devoiced high vowels, therefore, seem to be more likely to delete when the vowel is phonotactically predictable than when it is unpredictable.

PAC Number(s): 43.70.Fq, 43.70.Mn

I. INTRODUCTION

A. Background

The current study investigates the effects of recoverability—by way of phonotactic predictability—on the likelihood of vowel deletion as a consequence of the process of high vowel devoicing in Japanese. High vowel devoicing is considered to be an integral feature of standard modern Japanese (Imai, 2010), so much so that dictionaries exist with explicit instructions for devoicing environments (Kindaichi, 1995, pp.25–27). High vowel devoicing is typically described as involving phonemically short high vowels /i/ and /u/, which lose their phonation in C_1VC_2 sequences when the vowels are unaccented and both C_1 and C_2 are voiceless obstruents. For example, while the /u/ in /kúʃi/ 'free use' and /kuʃi/ 'skewer' are both between two voiceless obstruents, only /kuʃi/ 'skewer' undergoes devoicing because the vowel is unaccented. Likewise, the /u/ is unaccented in both /kuki/ 'stem' and /kuɡi/ 'nail', but only /kuki/ 'stem' undergoes devoicing because the /u/ is flanked by two voiceless stops. The likelihood of devoicing depends largely on the manner of the flanking consonants, where devoicing rates can be as low as 60% between two fricatives or between an affricate C_1 and a fricative C_2, but can be nearly 100% elsewhere (Maekawa and Kikuchi, 2005; Fujimoto, 2015). Although not the focus of this study, accented high vowels and non-high vowels can also devoice between voiceless obstruents but at much lower rates (<25%; Maekawa and Kikuchi, 2005), and unaccented high vowels optionally also devoice utterance finally after a voiceless fricative or affricate.

Despite the productivity of high vowel devoicing in Japanese and the amount of interest the phenomenon received in phonetics and phonology, there still is debate over whether the devoicing process results in only the loss of laryngeal adduction as the name suggests or can lead to complete deletion of the vowel through additional loss of the lingual and labial gestures associated with the vowel. The lack of consensus regarding how much of the vowel gestures is lost as part of the process stems in part from a lack of terminological, theoretical, and experimental consistency. Since there is disagreement on how much of the target high vowels is lost, the current study henceforth will use the term *unphonated* to refer to cases where phonation is lost but oral gestures of the vowel remain and *deleted* for cases when both phonation and oral gestures of the vowel are lost. The traditional term *devoicing* will be used to encompass both possibilities.

Theoretically, high vowel devoicing is assumed to be a postlexical process (Hirayama, 2009), which applies after lexical processes such as *rendaku*[1] (Ito and Mester, 2003) and structural processes such as syllabification and phonotactic evaluation (Boersma, 2009; Hayes, 1999; Zsiga, 2000). This is based on the observation that both underlying and epenthetic high vowels are targeted for devoicing, as exemplified by the CV sequence /ki/ in the Sino-Japanese compounds in (1) below. In (1a), the vowel /i/ is underlyingly present, wherease in (1b), the vowel is epenthetic (Ito, 1986; Ito and Mester, 2015; Kurisu, 2001; Tateishi, 1989). Because high vowel devoicing applies after phonotactic repairs, phonotactic constraints do not evaluate devoiced sequences, making both unphonated and deleted high vowels acceptable surface forms.

(1) a. |k̲i̲+tai| → /k̲i̲.tai/ → [k̲i̥̲tai] ‘expectation (time period+wait)’
b. |tek̲+tai| → /te.k̲i̲.tai/ → [tek̲i̥̲tai] ‘hostility (enemy+toward)’

This study aims to test the hypothesis that the choice between deletion and unphonating is dependent on the vowel’s recoverability (Varden, 2010). Recoverability refers to the ease of accessing the underlying form (i.e., stored mental representations) from a given surface form (i.e., actual, variable output signals; Mattingly, 1981; McCarthy, 1999; Chitoran et al., 2002), as in when accessing /kæt̲/ ‘cat’ from [kæt̲̚, kæt̲ʰ], for example. Recoverability comes largely from two sources: perceptibility of articulatory cues present in the acoustic signal or predictability based on linguistic knowledge, such as phonotactics. However, recoverability can be compromised if neither perceptibility nor predictability is sufficient. Varden (2010) states what seems to be a prevalent assumption in the Japanese high vowel devoicing literature, which is that since high vowels trigger allophonic variation on preceding /t, s, h/ (i.e., /t/ → [ʧi, ʦu]; /s/ → [ʃi, su]; /h/ → [çi, ɸu]), the underlying vowel is easily recoverable even if the vowel were to be phonetically deleted because the devoiced vowel is predictable in these contexts. For example, [ɸku] can only be analyzed as /huku/ ‘clothes’ because [ɸ_k] is a devoicing context, where the vowel to be recovered can only be one of /i, u/, and the mere presence of [ɸ] narrows the choice down to /u/ because [ɸ] can only occur as an allophone of /h/ preceding /u/. Because the context alone is sufficient for recovery, retaining oral gestures of the devoiced vowel to increase its perceptibility (e.g., [ɸu̥ku]) does little to improve recoverability. What Varden is proposing then is that a devoiced vowel is more likely to be deleted when phonotactic predictability is high, which also leads to the reverse prediction that a devoiced vowel is less likely to delete if phonotactic predictability is low.

A number of studies have proposed similar recoverability-conditioned coarticulation, where speakers seem to preserve or enhance the phonetic cues of a target segment in situations where the target segment would be less perceptible, such as when a phoneme inventory contains acoustically similar phonemes (Silverman, 1997) or in word-initial stop-stop sequences, where the closure of the second stop would obscure the burst of the first (Chitoran et al., 2002). However, whether C_1V coarticulation is similarly modulated by phonotactic predictability in Japanese has not been tested systematically.

B. Previous studies

There are primarily three ways in which devoiced high vowels are argued to be manifested acoustically: (i) by lengthening the burst/frication noise of C_1 (Han, 1994), (ii) by unphonating the vowel and coloring the C_1 burst/frication noise with the retained oral gestures without necessarily lengthening C_1 (Beckman and Shoji, 1984), and (iii) by deleting the vowel altogether (Vance, 2008). Each of the proposed manifestations has contradicting evidence in previous literature as discussed below.

Although it is commonly argued that C_1 is longer in devoiced syllables than in voiced syllables, the empirical evidence is not unanimous. Part of the problem in the lack of consensus regarding the effects of vowel devoicing on C_1 duration in Japanese is that there are differences in the methodologies and stimuli among the studies. While lengthening effects are reported for all consonant manners (Kondo, 1997), when no effect is found, it is generally studies that focus on fricatives. For example, Varden (1998) examines /k, t/ (where /t/ → [ʧi, ʦu]) and reports that the burst and aspiration of C_1 in devoiced syllables are significantly longer than the consonant portion of their corresponding voiced CV syllables. On the other hand, studies that focus on /s/ (→ [ʃi, su]; Beckman and Shoji, 1984; Faber and Vance, 2000) often report a lack of lengthening effect.

Additionally, studies that report lengthening effects generally assume that Japanese is mora-timed and that moras are roughly equal in duration. Based on these assumptions, the duration results of individual C_1 are often collapsed (Tsuchida, 1997; Nielsen, 2008), C_1 in devoiced contexts are compared to different segments in voiced contexts (Han, 1994), or the same segments from the same words that optionally devoice are compared to each other (Kondo, 1997). These practices are justified if moras in Japanese are indeed equal in duration, but as Warner and Arai (2001a,b) argue, the apparent rhythm in Japanese and the compensatory lengthening effect in relation to mora-timing might be epiphenomenal, stemming from a confluence of factors that result from the phonological structure of Japanese.

While it is conceptually plausible that the presence of an underlying vowel can be signaled solely by C_1 lengthening, especially if mora preservation is the reason behind it, much of the literature arguing for compensatory lengthening also reports formant-like structures, suggesting that the vowel is not completely deleted. A number of articulatory studies looking at /k, t, s/ as C_1 found that the glottis is wider when the vowel in a C_1VC_2 sequence is devoiced than when it is not, and that there is only one activity peak for the laryngeal muscles aligned with the onset of C_1 in devoiced sequences, resulting in a long frication or a frication-like burst release for stops (Fujimoto et al., 2002; Tsuchida et al., 1997; Yoshioka et al., 1980). Since there is no laryngeal activity associated with C_2 apart from the carry-over from C_1 and because the abduction peak for the glottis was found to be larger than the sum of two voiceless consonants, these results are interpreted to mean that the glottal gesture is being actively controlled to spread the feature [+spread glottis] from the first consonant to the second. As a consequence of this spreading, the intervening high vowel is devoiced. Despite the lack of a laryngeal gesture associated with phonation, presence of formant-like structures in the burst/frication noise of C_1 is often reported, which is taken as evidence of retained oral vowel gestures. For example, an acoustic study by

Varden (2010) reports visible formant structures apparent in the fricated burst noise of [ki̥, ku̥], which are interpreted to be the result of oral gestural overlap that allows consistent identification of the underlying devoiced vowel.

In contrast, Ogasawara (2013) reports a lack of visible formant structures in the burst/frication noise of /k, t/ in most devoiced cases and argues that this provides support for the claim that high vowel devoicing results in deletion rather than unphonating (Hirose, 1971; Yoshioka, 1981). The lack of apparent formant structures in the burst/frication noise of C_1, however, seems to be an inadequate criterion for measuring the presence of vocalic oral gestures. While Beckman and Shoji (1984) also report inconsistent presence of formant-like structures on the frication noise of /ʃ/, spectral measurements of [ʃ] showed a small yet noticeable influence of devoiced vowels on the aperiodic noise of the preceding fricative, where the mean frequency of [ʃu̥] was lower than [ʃi̥] by approximately 400 Hz, suggesting a coarticulatory effect of an unphonated vowel. Perceptually, this difference was enough to aid listeners in identifying the underlying vowel above the rate of chance (77% for [ʃi̥] and 67% for [ʃu̥]). Similar sensitivity to /ʃV/ coarticulation in Japanese listeners is also reported by Tsuchida (1994).

C. Predictability and coarticulation

The current study uses /ʧ, s, ç, ɸ/ as C_1 with high phonotactic predictability and /k, ʃ/ as C_1 with low phonotactic predictability. Although /ʃ, ʧ/ are more accurately alveopalatal consonants (i.e., /ɕ, ʨ/), the palatoalveolar symbols are used throughout the current study to make /ʃ/ more visually distinct from /ç/ and to make /ʧ/ consistent in place with /ʃ/. The bilabial stop /p/ is excluded because it rarely occurs word-initially, and the affricate [ʦ] is also excluded to keep the number of stimuli balanced between high and low predictability tokens.

There are two things to note regarding the chosen consonants. First, segments that were traditionally regarded as allophones are being used more phonemically in Japanese today. For example, although [ʧ] and [ɸ, ç] are allophones of /t, h/, respectively, before high vowels in native Japanese words, they are used phonemically in Sino-Japanese and loanwords. Minimal loan pairs such as [tiaː] ‘tier’ and [ʧiaː] ‘cheer’ show that [t, ʧ] can contrast on the surface before /i/, suggesting that words like ‘cheer’ contain an underlying /ʧ/ that surfaces faithfully, rather than an underlying /t/ that undergoes allophony. Additionally, /ɸ/ still neutralizes with /h/ before /u/, but /ɸ/ can precede every vowel of Japanese in loanwords (e.g., /ɸiɴ, ɸesu, ɸaʃʃoɴ, ɸoroː, ɸuriː/ ‘fin, fes(tival), fashion, follow(-up), free(lance)’). /ç/ also neutralizes with /h/ before /i/, but can precede all vowels except /e/ in both Sino-Japanese and loan words. Furthermore, /s/ is typically thought to neutralize with /ʃ/ before /i/, but as the predictability analysis below shows, [si] does occur on the surface, although it is still quite rare. Therefore, the current study regards /ʧ, s, ç, ɸ/ as phonemes that have extremely skewed phonotactic distributions that lead to higher levels of predictability.

Second, voiced and voiceless velar stops coarticulate with a following /i/ in Japanese (Maekawa, 2003; Maekawa and Kikuchi, 2005), as is often the case crosslinguistically. The question that remains unanswered, however, is whether the coarticulation leads to a categorical change of the consonants to neutralize with the phonemically palatalized velar stops of Japanese (e.g., /ki, kʲi/ → [kʲi]) or a relative fronting of the velar stops (e.g., /ki/ → [k̟i]). Spectral analyses have shown that the stop burst in /ki/ is significantly higher in frequency than /ku/ even in devoiced tokens (Kondo, 1997; Varden, 2010), suggesting either that velar fronting is categorical (i.e., /ki/ → [kʲi]) or that the underlying consonant is simply different (i.e, /kʲi/ vs. /ku/). However, perhaps due to the influence of Japanese orthography, the velar stops in [kʲi, ku] tend to be grouped together as /k/ when phonotactic distributions are calculated, making them distinct from the phonemically palatalized /kʲ/ as in /kʲa, kʲu, kʲo/ (Tamaoka and Makioka, 2004; Shaw and Kawahara, 2017). The current study follows the latter studies, grouping /ki, ku/ together for the purposes of calculating phonotactic predictability, but revisits this issue in Section IV after the acoustic results are analyzed.

1. Measuring predictability

Predictability is quantified using two Information-Theoretic (Shannon, 1948) measures: *surprisal*, which indicates how unexpected a vowel is after a given C_1, and *entropy*, which indicates the overall level of uncertainty in a given context due to competition amongst other possible vowels. If an unexpected vowel (high surprisal) occurs in an uncertain environment (high entropy), the vowel is difficult to predict. Conversely, a vowel with low surprisal occurring in a low entropy environment is easy to predict. Both measures are calculated based on the conditional probabilities of vowels after a given consonant, which can be written as $\Pr(v \mid C_1\_)$, which means the probability of vowel v occurring after consonant C_1. So for example, $\Pr(u \mid s\_)$ would be calculated as the frequency of /su/ divided by the frequency of /sV/ (any vowel after /s/).

Surprisal is the negative $\log_2$ probability. The log tranform turns the probability into bits, which indicates the amount of information (or effort) necessary to predict a vowel. The equation for surprisal is given below.

$$\text{Surprisal: } -\log_2 \Pr(v \mid C_1\_)$$

Entropy is the weighted average of surprisal in a given context. The untransformed probability of vowel v in context $C_1\_$ serves as the weight for the surprisal of the same vowel and context. The equation for entropy calculations is given below.

$$\text{Entropy (H): } \sum \Pr(v \mid C_1\_) * (-\log_2 \Pr(v \mid C_1\_))$$

When given a C_1C_2 sequence with no apparent intervening vowel, experience with high vowel devoicing informs the Japanese listener that the most likely candidates for vowel recovery must be /i, u/ because non-high vowels and long vowels typically do not devoice. There is no upper bound to surprisal, but the theoretical maximum of entropy (highest uncertainty) in any given consonantal context with two possible vowels is 1.000 ($-\log_2 p(0.5)$), where both vowels occur with equal probabilities (1/2 = 0.5).

Below in Table I are entropy and surprisal measures calculated from the “Core” subset of the Corpus of Spontaneous Japanese (Maekawa, 2003; Maekawa and Kikuchi, 2005) for the consonants included in the current study.

TABLE I: C_1 consonants used in stimuli with overall entropy and surprisal of /i, u/. Ordered from highest to lowest entropy.

	IPA	*Entropy*	*Surprisal /i/*	*Surprisal /u/*
low predictability	k	9.998e-01	0.979	1.021
	ʃ	0.555	0.199	2.955
high predictability	ɸ	0.123	5.903	0.024
	s	0.042	7.762	0.007
	ʧ	0.013	0.002	9.768
	ç	0.008	0.001	10.653

None of the entropy and surprisal values are at zero across all environments, meaning both /i, u/ occur after each C_1. However, there are notable differences between /k, ʃ/ and /ɸ, s, ʧ, ç/. First, the entropy is near-zero for /ɸ, s, ʧ, ç/, which means that given any of these C_1, there is essentially no uncertainty regarding the vowel that will follow. This is not true for /k, ʃ/, however, where entropy is closer to the maximum of 1.000 than to the minimum of 0.000. Second, surprisal values for /u/ following /ɸ, s/ and for /i/ following /ʧ, ç/ are also near-zero because the high vowels occur with frequencies greater than 0.980. While there are differences between /i, u/ surprisal values in the /k, ʃ/ contexts as well, the differences are not as large. In the case of /k/, /i, u/ have approximately the same relative frequencies (0.507 vs. 0.493, respectively), and while /i/ is the more frequent vowel after /ʃ/, /u/ still occurs with a non-negligible frequency of 0.129. Together, the entropy and surprisal calculations show that devoiced high vowels can be predicted with near-absolute certainty after /ɸ, s, ʧ, ç/ but not after /k, ʃ/.

2. Possible effects of predictability on coarticulation

There are three main possibilities with respect to the question of how predictability affects devoiced vowels. The first is that high vowel devoicing is blind to predictability and is driven primarily by Japanese phonotactics, which has a strict CVCV structure that disallows tautosyllabic clusters (Kubozono, 2015). If this is the case, then no difference between low predictability and high predictability C_1 would be found, where the devoiced vowel does not delete but becomes unphonated instead, coloring the burst or frication noise of C_1 to signal the presence of the target vowel (Beckman and Shoji, 1984; Varden, 2010). The second is that the choice between deletion and unphonating is not systematic but rather a consequence of how the devoiced vowel happened to be lexicalized for the speaker. Ogasawara and Warner (2009) found in a lexical judgment task that when Japanese listeners were presented with voiced forms of words where devoicing is typically expected, reaction times were longer than when presented with devoiced forms. This suggests that devoiced forms, despite their phonotactic violations, can have a facilitatory effect on lexical access due to their commonness, making vowel recovery unnecessary (Cutler et al., 2009; Ogasawara, 2013). The third and last option, which this study proposes, is that high vowel devoicing is constrained by recoverability. In this case, the presence of the devoiced vowel would be observable either by lengthening or spectral changes of C_1 burst/frication when the predictability of the target vowel is unreliable from a given C_1 to aid recovery from the coarticulatory cues as in the case of /k, ʃ/, but not when predictability is high, as in the case of /ʧ, s, ɸ, ç/. This last outcome would also be compatible with the idea that devoiced forms are lexicalized as such (Ogasawara and Warner, 2009), but with the caveat that whether the vowel is unphonated or deleted is dependent on predictability from context.

While this study does not explore sociolinguistic factors that affect high vowel devoicing, it is worth noting that men have been reported to devoice more than women (Okamoto, 1995) and that devoicing rates are higher overall in younger speakers (Varden and Sato, 1996). However, Imai (2010) found that while younger speakers did tend to devoice more, this was only true for men. Young female speakers were actually shown to devoice the least among all age groups. Based on these findings, Imai proposes that high vowel devoicing might be being utilized actively as a feature of gendered speech. If high vowel devoicing is being utilized as a sociolinguistic feature, then the process could not be purely phonological or phonetic, and thus a balanced number of men and women were recruited to investigate any gender-based differences.

II. MATERIALS AND METHODS

A. Participants

Twenty-two monolingual Japanese speakers (12 women and 10 men) were recruited in

Tokyo, Japan. All participants were undergraduate students born and raised in the greater Tokyo area and were between the ages 18 and 24. Although all participants learned English as a second language as part of their compulsory education, none had resided outside of Japan for more than six months and have not been overseas within a year prior to the experiment. All participants were compensated for their time.

B. Materials

The stimuli for the experiment were 160 native Japanese and Sino-Japanese words with an initial C_1iC_2 or C_1uC_2 target sequence. The stimuli were controlled to be of medium frequency (20 to 100 occurrences, which is the mean and one standard deviation from the mean, respectively) based on the frequency counts from a corpus of Japanese blogs (Sharoff, 2008). Any gaps in the data were filled with words of comparable frequency based on search hits in Google Japan (10 million to 250 million). Since high vowel devoicing typically occurs in unaccented syllables, an accent dictionary of standard Japanese (Kindaichi, 1995) was used as reference to ensure that none of the stimuli had a target vowel in an accented syllable.

The stimuli were divided into *low predictability* and *high predictability* groups as discussed above. Since only high vowels are systematically targeted for devoicing and recovery, predictability refers specifically to the predictability of backness of high vowels. Examples of devoicing stimuli are shown in Table II below.

TABLE II: Example of devoicing stimuli by C_1 and vowel.

stimulus type	C_1	*V*	*example*	*gloss*
low predictability	k	i	kikai	‘chance’
		u	kuki	‘stalk’
	ʃ	i	ʃitagi	‘underwear’
		u	ʃutokeɴ	‘capital area’
high predictability	tʃ	i	tʃikʲuː	‘earth’
	s	u	sukui	‘help’
	ɸ	u	ɸukoː	‘unhappiness’
	ç	i	çiteː	‘denial’

As shown above, for the low predictability group, C_1 was either /k, ʃ/ after which both /i, u/ can occur. For the high predictability group, C_1 was one of /ʧ, s, ɸ, ç/, after which only one of the high vowels is likely. The two groups were further divided into *devoicing* and *voicing* contexts. The difference between devoicing and voicing tokens was that C_2 was always a voiceless stop for devoicing contexts as shown above, but a voiced stop for voicing tokens. Since high vowel devoicing typically requires the target vowel to be flanked by two voiceless obstruents, it was expected that devoicing would not occur in the voicing contexts. The C_1 and C_2 combinations resulted in fricative-stop, affricate-stop, or stop-stop contexts. These contexts were chosen for two reasons: (i) these are contexts in which the loss of phonation in high vowels is reported to occur systematically and categorically (Fujimoto, 2015), and (ii) the C_2 stop closure clearly marks where the previous segment ends. There were 10 tokens per C_1V combination within each context, for a total of 160 tokens (80 devoicing and 80 voicing).[2]

C. Design and procedure

All tokens were placed in the context of unique and meaningful carrier sentences of varying lengths. Most carrier sentences were part of a larger story, and thus no two carrier sentences were identical. All carrier sentences contained at least one stimulus item, and the sentences were constructed so that no major phrasal boundaries immediately preceded or followed the syllable containing the target vowel. An example carrier sentence, which was actually uttered by a weather forecaster in Japan, is given below with glosses.

(2) manatsu-no ʃigaiseɴ-ni-wa ki-o-tsuke-maʃoː
midsummer's ultraviolet rays-DAT-TOP be careful.VOL
'Let's be careful of midsummer's ultraviolet rays'

DAT = dative; TOP = topic; VOL = volition

The carrier sentences were presented one at a time to the participants on a computer monitor as a slideshow presentation. The participants advanced the slideshow manually, giving the participants time to familiarize themselves with the sentences. They were also allowed to take as many breaks as they thought was necessary during the recording. All instructions were given in Japanese, and participants were prompted to repeat any sentences that were produced disfluently.

All participants were recorded in a sound-attenuated booth with an Audio-Technica ATM98 microphone attached to a Marantz PMD-670 digital recorder at a sampling rate of 44.1 kHz at a 16 bit quantization level. The microphone was secured on a table-top stand, placed 3-5 inches from the mouth of the participant.

D. Data Analysis

Once the participants were recorded, the waveform and spectrogram of each participant were examined in Praat to (a) code each token for devoicing, (b) to measure the duration of C_1 and the following vowel, and (c) to measure the center of gravity of C_1 burst/frication noise. The spectrogram settings were as follows: pre-emphasis was set at +6 dB, dynamic range was set at 60 dB, and autoscaling was turned off for consistency of visual detail. Because visual inspection alone is an inadequate method for determining the presence of vowel coarticulation on C_1 (Beckman and Shoji, 1984), tokens were simply coded for “devoicing”, a term used to collectively refer to unphonating and deletion of the vowel. The criteria used for devoicing status are described in the following section.

1. Devoicing analysis

Vowels in devoicing environments were coded as voiced if there was phonation accompanied by formant structures between C_1 and C_2. Vowels were coded as devoiced when there was no phonation between C_1 and C_2. Below in Figure 1 are examples from the same female speaker. On the left is a voiced vowel in the word [kuki] ‘stalk’, which shows clear phonation and formant structures between C_1 and C_2. On the right is a devoiced vowel in the word [ku̥teɴ] ‘period’, where there is neither phonation nor formant structures between C_1 and C_2.

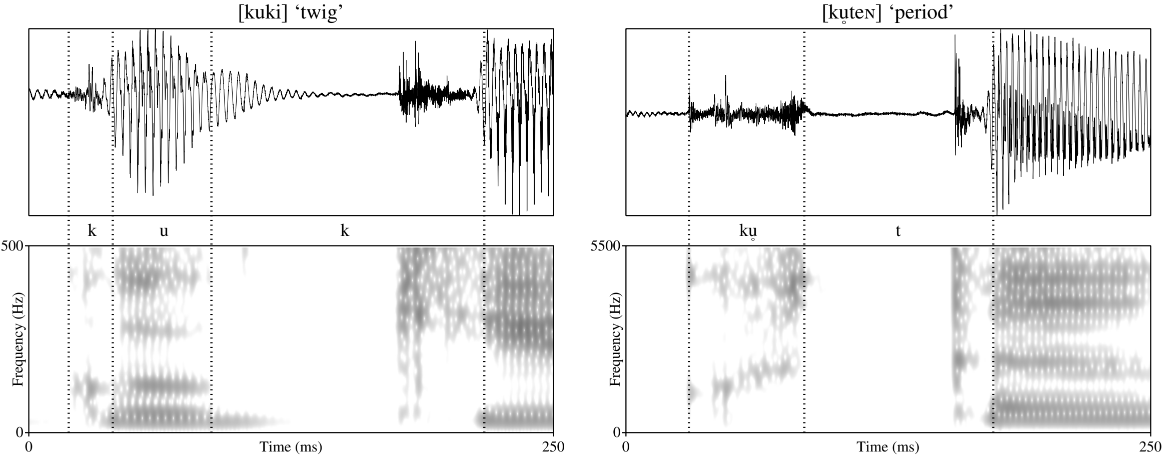

FIG 1: Waveform and spectrogram of voiced (left) and devoiced (right) vowels in devoicing environments, showing landmarks for C_1, vowel, and C_2 duration.

The coding criteria were similar for voicing tokens. Vowels were coded as voiced if phonation and formant structure were both present between C_1 and C_2. Otherwise, vowels were coded as devoiced. Below in Figure 2 are examples from another female speaker. On the left is a voiced vowel in the word [ʃɯgeː] ‘handicraft’, where there is a clear formant structure accompanying phonation. On the right is a rare case of a devoiced vowel in a voicing word [ʃɯ̥daika] ‘theme song’, where there is low frequency pre-voicing preceding C_2 but no formant structure.

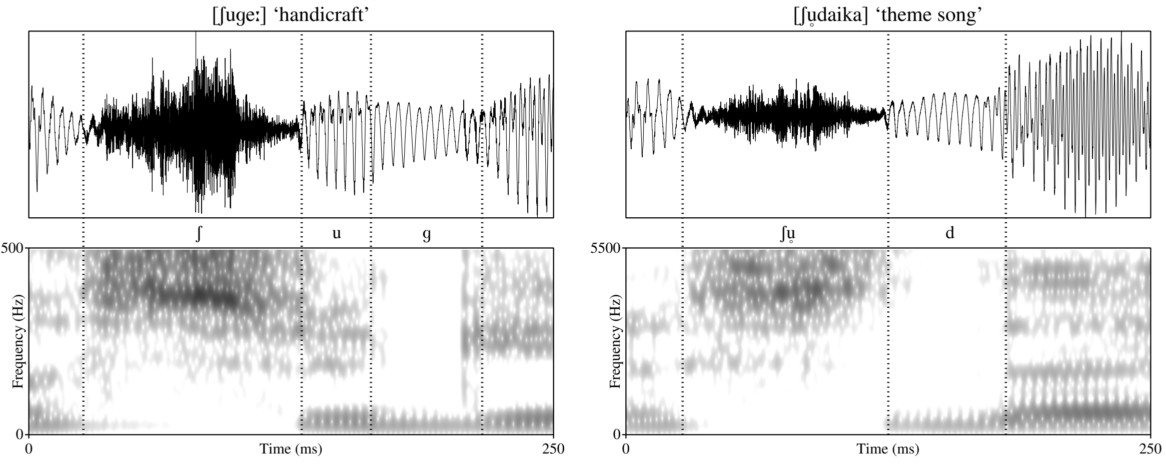

FIG 2: Waveform and spectrogram of voiced (left) and devoiced (right) vowels in voicing environments, showing landmarks for C_1, vowel, and C_2 duration.

2. Duration analysis

Once all tokens were coded for devoicing status, duration measurements were taken to investigate how devoicing affects the gestural timing of C_1 and the target high vowel. For [k] and [ʧ], duration measurements excluded the silence from closure. For [k], measurements included only the aperiodic burst energy, and for [ʧ], the burst and frication noise. For fricative C_1, duration measurements included the entire aperiodic frication noise. For tokens coded as devoiced, C_1 measurements were assumed to include the devoiced vowel because the vowel could not be isolated from C_1 reliably. For voiced tokens, C_1 was measured from the onset of burst/frication noise to the onset of vowel F2. For both duration and center of gravity analyses, only devoiced tokens in devoicing environments and voiced tokens in voicing environments were included.

3. Center of gravity analysis

Center of gravity (COG), which is the amplitude weighted mean of frequencies present in the signal (Forrest et al., 1988), was also calculated for C_1 to investigate the presence of coarticulation between C_1 and the target vowel. COG measurements are used based on Tsuchida (1994), who found that Japanese listeners rely primarily on C_1 centroid frequency (i.e., COG) to identify devoiced vowels. COG measurements are known to be particularly sensitive to changes in the front oral cavity (Nittrouer et al., 1989), so the effects of coarticulation between a vowel and C_1 on COG values are expected to differ by the backness and roundedness of the vowel as well as C_1 place of articulation. The predicted effects of vowel coarticulation on each C_1 are discussed in detail in Section III. C together with the results.

Before measuring COG values, the sound files were high pass filtered at 400 Hz to mitigate the effects of f0 on the burst/frication noise. The filtered sound files were then down-sampled to 22,000 Hz. The COG values measured therefore were taken from FFT spectra in the band of 400 to 11,000 Hz (Forrest et al., 1988; Hamann and Sennema, 2005). With the exception of /k/, two center of gravity (COG) measurements were taken from 20 ms windows for each C_1: one starting 10 ms after the beginning of C_1 burst/frication (COG1) and one ending 10 ms before the end of C_1 burst/frication (COG2). The 10 ms buffers were used to mitigate the coarticulatory effects of segments immediately adjacent to C_1. For /k/, COG measurements were taken from a single 20 ms window at the midpoint of the burst. Two COG measurements could not be taken from /k/ because /k/ durations in voiced tokens were too short for two measurements. /k/ tokens shorter than 20 ms were excluded from analysis, which resulted in the loss of five tokens, or 0.6% of the /k/ data. Since the vocalic gesture of the following vowel most likely begins during the stop closure for /k/ (Browman and Goldstein, 1992; Fowler and Saltzman, 1993), the single COG measurement is assumed to be equivalent to the COG2 measurements of other consonants. Voiced tokens provide the baseline C_1V coarticulation, and comparing the COG1 and COG2 values of devoiced tokens to those of voiced tokens allows for testing of whether coarticulatory effects that are comparable to voiced tokens are present in devoiced tokens at the beginning and end of C_1.

III. RESULTS

Statistical analyses were performed by fitting linear mixed effects models using the *lme4* package (Bates et al., 2015) for R (R Core Team, 2017). In order to identify the maximal random effects structure justified by the data, a model with a full fixed effects structure (i.e., with interactions for all the fixed effects) and the most complex random effects structure was fit first (Barr et al., 2013). If the model did not converge, the random effects structure was simplified until convergence was reached while keeping the fixed effects constant. The simplest random effects structure considered was one with random intercepts for participant and word with no random slopes.

Once the maximal random effects structure was identified, a Chi-square test of the log likelihood ratios was performed to identify the best combination of fixed effects. A complex model with all interaction terms was fit first, which was then gradually simplified by removing predictors that did not significantly improve the fit of the model, starting with interaction terms. The simplest model considered was a model with no fixed effects and only an intercept term.

A. Devoicing rate

Devoicing rates were at or near 100% in environments where devoicing was expected, which confirms that loss of phonation in these contexts is phonological. Devoicing rates were less than 25% in environments where devoicing was not expected. This is shown in Table III below.

TABLE III: Devoicing rate by C_1V and context.

stimulus type	C_1	*V*	*devoicing*	*voicing*
low predictability	k	i	1.000	0.077
		u	0.959	0.032
	ʃ	i	1.000	0.086
		u	0.973	0.073
high predictability	ʧ	i	1.000	0.191
	ç	i	1.000	0.015
	ɸ	u	1.000	0.042
	s	u	1.000	0.214
overall			0.992	0.091

A mixed logit model was fit using the *glmer()* function of the *lme4* package for the overall devoicing rate with context, predictability, gender, and their interactions as predictors. Vowel was not included as a predictor because it is redundant for high predictability tokens since only one vowel is allowed. Random intercepts for participant and word were added to the model. By-participant random slopes for context and predictability as well as by-word random slopes for gender were also included in the model. The final model retained the full random effects structure. The following predictors were removed from the fixed effects structure of the final model as they were not significant contributors to the fit of the model: three-way interaction (p = 0.999), context:gender interaction (p = 0.902), and predictability:gender interaction (p = 0.062). The function for the final model, therefore, was as follows:

model = glmer(devoicing ∼ context + predictability + gender + context:predictability + (1 + context + predictability | participant) + (1 + gender | word), family = binomial(link = 'logit'), data = non-loanwords)

The results of the final model showed that the difference in devoicing rates between devoicing and voicing contexts was significant ($p < 0.001$) and that men were more likely to devoice than women (p = 0.018). Predictability and the context:predictability interaction did not have significant effects (p = 0.237 and 0.724, respectively).

An additional analysis was performed on just the voicing subset of the data because vowels in devoicing contexts devoiced essentially 100% of the time and had no between-participant differences to test statistically. First, a mixed logit model was fit to the low predictability voicing tokens with gender, C_1, vowel, and their interactions as predictors. Random intercepts for participant and word were included in the model. By-participant random slopes for C_1 and vowel, and by-word random slopes for gender were also included. /ʃ/ tokens as produced by female participants were the baseline. However, none of the predictors were significant contributors to the fit of the model, and a Chi-square test showed the fit of the intercept-only model was not significantly different from more complex models. In other words, /k, ʃ/ had similar devoicing rates in voicing contexts regardless of vowel or gender.

Second, a mixed logit model was fit to the high predictability voicing tokens with gender, C_1, and their interaction as predictors. Random intercepts for participant and word were included in the model. By-participant random slopes for C_1 and by-word random slopes for gender were also included. The interaction term was not a significant contributor to the model (p = 0.078), and thus was removed from the final model. /ʧ/ tokens as produced by female participants were the baseline. The results showed that male participants were more likely to devoice than women (p =

0.012). C_1 did not have a significant effect ($p = 0.171$, 0.092, and 0.517 for /ɸ, ç, s/ respectively). The separate analyses of voicing tokens suggest that male participants devoice more in high-predictability environments, where devoicing is not actually phonologically conditioned (e.g., /ɸugoːri/ → [ɸu̥goːri] 'unreasonable').

B. Duration

Previous studies that report lengthening effects of devoicing on C_1 generally have focused on /k, t/ (Varden, 1998), while studies that report a lack of such effect focused on /s, ʃ/ (Beckman and Shoji, 1984; Vance, 2008). There are two confounded differences between /k, t/ and /s, ʃ/ that may be contributing to the contrary results: manner and inherent duration. /k, t/ are non-continuants while /s, ʃ/ are continuants, but it is also the case that /k/ burst and /ʧ/ burst/frication are inherently much shorter than the frication noise of /s, ʃ/. This means that the contrary results could be due to either or both of these differences. /ɸ, ç/ are therefore crucial in teasing apart the two factors because /ɸ, ç/ are fricatives but are also similar in duration to the frication portion of /ʧ/ in Japanese.[3]

Duration results are shown in Figure 3 below. The results suggest that overall C_1 burst/frication durations are not different between women and men. Devoicing has a lengthening effect only on non-fricative obstruents (i.e., /ki, ku, ʧi/). For fricatives, devoicing has either no effect (i.e., /ɸu/) or a shortening effect (i.e., /çi, su, ʃu, ʃi/).

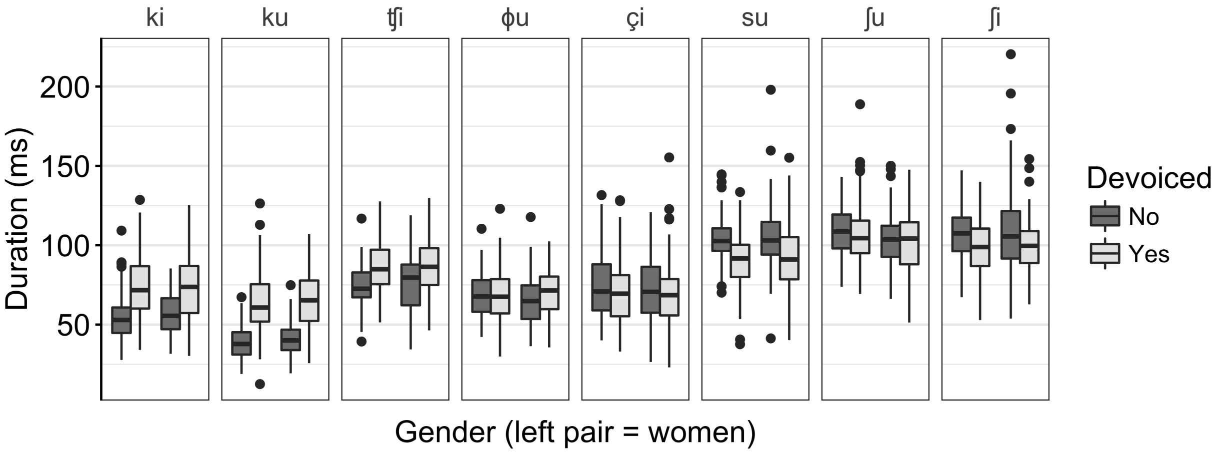

FIG 3: C_1 duration in ms by C_1V, gender, and devoicing.

A linear mixed effects regression model was fit to the overall duration results with devoicing, gender, C_1, and their interactions as predictors. Again, vowel was not included as a predictor because it is only meaningful for /k, ʃ/ tokens. Random intercepts for participant and word were added to the model. By-participant random slopes for context and C_1 were also included in the model, as well as by-word random slopes for gender. *p*-values were calculated by using the *lmerTest* package (Kuznetsova et al., 2017) for R.

The final model retained the full random effects structure. The following non-significant predictors were removed from the final model: three-way interaction ($p = 0.304$), devoiced:gender interaction ($p = 0.927$), gender:C_1 interaction ($p = 0.608$), and gender ($p = 0.580$). The final model therefore retained devoicing, C_1, and their interaction as predictors. The function for the final model was as follows:

*model = lmer(duration ∼ context * C1 + (1 + context + C1 | participant) + (1 + gender | word), control=lmerControl(optimizer=“bobyqa”), REML = F, data = non-loanwords)*

The final model’s results are summarized below in Table IV. Voiced /k/ tokens are the baseline.

TABLE IV: Linear mixed effects regression model results for overall C_1 duration.

	ms	S.E.	*t*	
(Intercept)	47.365	2.264	20.917	***
devoiced	22.068	3.106	7.106	***
ɸ	20.464	3.516	5.819	***
ç	26.808	3.746	7.156	***
ʧ	27.399	3.634	7.539	***
s	55.317	3.751	14.749	***
ʃ	59.454	3.155	18.844	***
devoiced:ɸ	-20.396	4.877	-4.182	***
devoiced:ç	-25.340	4.964	-5.105	***
devoiced:ʧ	-10.514	4.895	-2.148	*
devoiced:s	-33.451	4.903	-6.823	***
devoiced:ʃ	-27.009	3.983	-6.781	***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

The results show that devoicing indeed has a lengthening effect of 22 ms on /k/. The intercept estimates for C_1 predictors show that all other C_1 are significantly longer than the /k/ baseline. The negative values of the estimates for the devoiced:C_1 interaction predictors also show that devoicing has a smaller lengthening effect on all other C_1 relative to the /k/ baseline.

The model above only shows how other C_1 differ from /k/. In order to explore whether devoicing actually had significant effects on individual C_1, differences of least squares means were calculated from the final model using the *difflsmeans()* function of the *lmerTest* package (Kuznetsova et al., 2017). The results showed that devoicing had a significant lengthening effect on /ʧ/ (11.6 ms, $p = 0.007$). The fricatives on the other hand showed varying effects. Devoicing had a non-significant lengthening effect of 1.7 ms on /ɸ/ ($p = 0.691$) and non-significant shortening effects of 3.3 ms on /ç/ ($p = 0.447$) and 4.9 ms on /ʃ/ ($p = 0.114$). However, devoicing had a significant shortening effect of 11.4 ms on /s/ ($p = 0.008$).

A separate linear mixed effects regression model was fit to low predictability tokens (i.e., /k, ʃ/) to investigate the effects of vowel type. Since the overall model above already showed that devoicing had a lengthening effect on /k/, the baseline was set to /ʃ/. Devoicing status, C_1, vowel type, and their interactions were included as predictors. Random intercepts by participant and word were included. By-participant random slopes for devoicing, C_1, and vowel type were also included, as well as by-word random slopes for gender. The final model retained the full random effects structure. The three-way interaction term and devoicing:vowel interaction were not significant contributors to the model ($p = 0.755$ and 0.126, respectively) and were removed from the fixed effects structure of the final model.

The results of the final model showed that although devoicing had a slight shortening effect of 5 ms and the vowel /u/ had a slight lengthening effect of 3 ms on /ʃ/, neither was significant ($p = 0.131$ and 0.285, respectively). Also, as was shown in the overall model above, devoicing had a significant lengthening effect of 22 ms on /k/ ($p < 0.001$).

C. Center of gravity (COG)

As discussed in Section II. D. 3, two COG values were measured for each C_1 using a 20 ms window, one beginning 10 ms after the start of C_1 (COG1), and one ending 10 ms before the end of C_1 (COG2). /k/ tokens were the exception, where only one COG value was measured using a 20 ms window centered at the middle of the burst, because /k/ bursts were too short. The single COG measurement of /k/ is considered to be equivalent to the COG2 measurements of other consonants for the purposes of statistical analysis, since it measures the end of the segment.

COG is sensitive primarily to changes in the front cavity (Nittrouer et al., 1989) but also constriction strength (Hamann and Sennema, 2005; Kiss and Bárkányi, 2006). In general, C_1V coarticulation is expected to lower the COG of C_1 but for different reasons. Although the high back vowel of Japanese has traditionally been regarded as unrounded (i.e., [ɯ]), a recent articulatory study by Nogita et al. (2013) showed that the high back vowel is actually closer to a rounded high central vowel [ʉ] in younger speakers. So for /ʃ/, /u/ coarticulation is expected to result in lower COG than /i/ coarticulation due to lip rounding, which would increase the size of the front oral cavity. /i/ coarticulation is also expected to lower COG, as the tongue shifts back towards the palate. The effects of coarticulation for /ʧ/ should be similar to /ʃi/, where lingual movement towards the palate for /i/ would increase the front cavity size and lower COG. For /s/, coarticulation with /u/ should lead to lower COG as a result of lip protrusion and the tongue shifting back. Because /ɸ, ç/ are essentially identical in place with the vowels that can devoice after them, changes in COG are expected to come primarily from constriction strength rather than change in the length of the front oral cavity[4], where weakening constriction lowers the amplitude of the higher frequencies and results in a lower COG value overall (Hamann and Sennema, 2005; Kiss and Bárkányi, 2006). In other words, for /ɸ/, coarticulation with /u/ would result in more lip rounding and weaker constriction, both contributing to lower COG values. For /ç/, coarticulation with /i/ would make the fricative more vowel-like with a weaker constriction, also resulting in lower COG values.

Given the expected lowering effect of C_1V coarticulation overall, there are three possible effects of devoicing. First, if a devoiced vowel is simply unphonated, where only phonation is lost and the oral gestures associated with the vowel are retained, devoiced tokens should show similar COG values as voiced tokens. Second, devoicing may show increased coarticulation between C_1 and the target vowel to aid the perceptibility of the target vowel, resulting in lower COG values for devoiced tokens than for voiced tokens (Tsuchida, 1994). Third, the vowel could delete as a consequence of devoicing, and since there is no intervening vowel target, this would allow coarticulation with the following consonant (Shaw and Kawahara, 2018; Tsuchida, 1994), which would be most apparent towards the end of C_1 (i.e., COG2). Since COG is affected by the size of the front oral cavity and constriction strength, the effects of deletion would depend on the place of C_2, which was either /k, t/ for devoicing tokens. Generally, for alveolar and alveopalatal C_1 (i.e., /s/ and /ʃ, ʧ/), coarticulation with /t/ would lead to higher COG values as the tongue shifts forward and constriction strength increases, while coarticulation with /k/ would lead to lower COG values as the tongue shifts back towards the palate. For /ç, k/, coarticulation with either C_2 would raise COG – /t/ due to tongue shifting forward and /k/ due to strengthening constriction. For /ɸ/, devoicing is expected to raise COG2 due to stronger labial constriction, unaffected by C_2 place. Since C_1k coarticulation can sometimes lower COG much like C_1V coarticulation, the COG analyses below focus on stimuli with alveolar C_2, so that C_1V coarticulation, which would lower COG, and C_1t coarticulation, which would raise COG, can be easily distinguished.

1. COG1 results and analysis

A summary table for all COG results is provided in Section IV (Table VII). Shown below in Figure 4 are COG1 results. C_1 /k/ is excluded, since there is only one COG measure for the consonant, which is regarded as equivalent to COG2 of other C_1. The figure suggests that devoicing has a lowering effect on COG1 for both /ʃi, ʃu/ for women but only for /ʃi/ for men. Devoicing also seems to have a raising effect for /çi, ɸu/. /ʧ, s/ do not show any effect of devoicing.

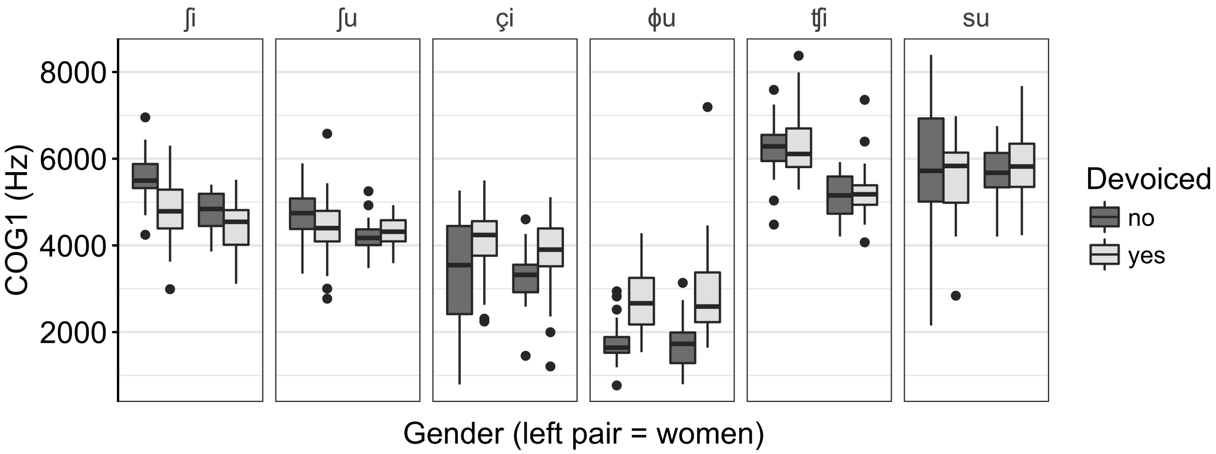

FIG 4: COG1 in Hz by C_1V, gender, and devoicing.

A model with the following structure was fit initially to the data:

model = lmer(COG1/2 $\sim$ *devoicing ** C_1 ** gender + (1 + devoicing* | *participant) + (1 + gender* | *word), control=lmerControl(optimizer=”bobyqa”), REML = F, data = alveolar* C_2*)*

As was the case for duration analyses, vowel was not included as a predictor since it is only relevant for /ʃ, k/. The final model excluded the following non-significant predictors: three-way interaction ($p = 0.243$) and devoicing:gender ($p = 0.163$). The results of the final model are presented in Table V below. Voiced /ɸu/ tokens as produced by female speakers are the baseline.

TABLE V: Linear mixed effects regression results: COG1 (excludes C_1 /k/).

	Hz	S.E.	*t*	
(Intercept)	1770	186.90	9.473	***
devoiced	1153	203.23	5.672	***
male	-54	191.55	-0.283	
ç	1567	257.86	6.075	***
s	4027	234.85	17.148	***
ʧ	4376	232.94	18.788	***
ʃ	3154	201.57	15.647	***
devoiced:ç	-458	293.70	-1.560	
devoiced:s	-1165	307.51	-3.790	***
devoiced:ʧ	-998	275.69	-3.621	***
devoiced:ʃ	-1314	247.63	-5.308	***
male:ç	-138	180.32	-0.764	
male:s	-2	180.95	-0.012	
male:ʧ	-1036	184.14	-5.625	***
male:ʃ	-391	146.37	-2.673	**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

The results show that for /ɸ/, devoicing has a significant raising effect for both men and women. Since the model above only shows how other C_1 compare to /ɸ/, differences of least squares means were calculated for a more detailed investigation into the other consonants. For /ç/, devoicing was also shown to have a significant raising effect of 695 Hz (*p* = 0.002), and there were no gender effects. For /s/, neither devoicing nor gender had a significant effect. For /ʧ/, devoicing had no significant effect but male speakers had significantly lower COG1 (-1090 Hz; *p* < 0.001). The overall model showed that male speakers also had lower COG1 for /ʃ/, but since the results collapse the two possible vowels after /ʃ/, a separate model was fit to test for vowel-specific effects.

The initial model for /ʃ/ tokens was as follows:

*model = lmer(COG1 ∼ devoicing * vowel * gender+ (1 + devoicing * vowel | participant) + (1 + gender | word), control = lmerControl(optimizer = 'bobyqa'), REML = F, data = alveolar* C_2*)*

The final model retained the full random effect structure, but excluded the three-way (*p* = 0.883) and vowel:gender ($p = 0.089$) interaction terms from its fixed effects structure. Male speakers were shown to have lower COG1 by 644 Hz ($p < 0.001$). The model also showed that /u/ had significant lowering effects of 687 Hz on voiced tokens ($p < 0.001$) and 289 Hz on devoiced tokens ($p = 0.035$), suggesting that coarticulation with /u/ is evident from the very beginning of the consonant, making the contrast between /ʃi/ and /ʃu/ more salient. Additionally, devoicing had a significant lowering effect of 531 Hz on /ʃi/ tokens produced by female speakers ($p = 0.001$), which suggests that there is an effort to make the identity of the devoiced /i/ vowel perceptually more salient by increasing the CV overlap. However, the lowering effect was not significant for /ʃu/ tokens (-132 Hz; $p = 0.233$), and male speakers showed no effect of devoicing (-96 Hz; $p = 0.379$).

2. COG2 results and analysis

COG2 results are shown in Figure 5 below, where devoicing seems to have a raising effect on the COG2 of all consonants.

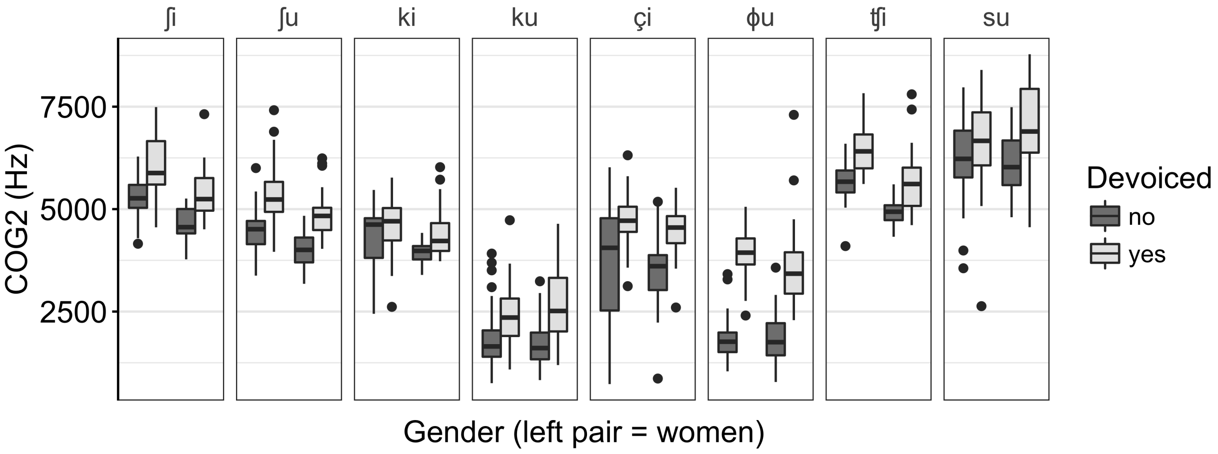

FIG 5: COG2 in Hz by C_1V, gender, and devoicing.

The same full linear mixed effects regression model used for COG1 was fit to COG2 initially. The final model excluded the three way ($p = 0.151$), devoicing:gender ($p = 0.398$), and devoicing:C_1 ($p = 0.358$) interaction terms. The results of the model are presented in Table VI below. Voiced /ɸu/ tokens as produced by female speakers are the baseline.

TABLE VI: Linear mixed effects regression results: COG2 (all C_1).

	Hz	S.E.	*t*	
(Intercept)	2427	263.69	9.205	***
devoiced	1031	143.44	7.186	***
male	-230	156.34	-1.470	
ç	1217	341.30	3.567	***
s	3591	366.34	9.802	***
ʧ	3029	334.35	9.059	***
ʃ	2322	301.80	7.695	***
k	-52	292.87	-0.176	
male:ç	-31	167.53	-0.186	
male:s	305	182.78	1.671	$^{\cdot}$
male:ʧ	-540	171.63	-3.144	**
male:ʃ	-348	143.88	-2.416	*
male:k	177	136.85	1.297	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

COG2 results largely mirror those of COG1, where devoicing has a raising effect for /ɸ/, and male speakers have significantly lower COG values for /ʃ, ʧ/. The fact that devoicing:C_1 interaction is not a significant predictor means that the raising effect of devoicing is evident across all C_1.

COG1 analysis showed that devoicing had a lowering effect on /ʃ/, although the effect was significant only in /ʃi/ tokens produced by female speakers. The general model above, however, suggests that devoicing could have a raising effect on COG2 instead, perhaps due to coarticulation with C_2 (Tsuchida, 1994). A separate model was fit to /ʃ/ tokens to test for effects of vowel type on COG2. The full model had the same structure as the model fit to COG1 data, and the final model for /ʃ/ COG2 retained the full random effects structure but only devoicing, vowel, and gender as predictors. Three-way interaction ($p = 0.399$), gender:vowel ($p = 0.939$), devoicing:vowel ($p = 0.710$), and devoicing:gender ($p = 0.145$) were non-significant predictors and removed from the final model. /u/ had a significant lowering effect of 680 Hz ($p < 0.001$) showing that the lowering effect observed in COG1 is retained throughout the consonant. Male speakers were also shown to have lower COG2 by 542 Hz ($p = 0.001$). Devoicing had a significant raising effect of 807 Hz ($p < 0.001$), suggesting coarticulation with C_2.

A separate model was also fit to /k/ tokens to test for vowel effects. The final model for /k/ retained the full random effects structure and only vowel and devoicing as predictors. Three-way interaction (p = 0.491), vowel:devoicing (p = 0.195), devoicing:gender (p = 0.157), vowel:gender (p = 0.241), and gender (p = 0.775) were non-significant predictors and removed from the final model. /u/ had a lowering effect of 2166 Hz ($p < 0.001$) and devoicing had a raising effect of 560 Hz ($p < 0.001$).

IV. DISCUSSION

The aim of this study was to investigate the acoustic properties of high vowel devoicing in Japanese—specifically, what cues in the signal allow the recovery of a devoiced vowel and whether gender and phonotactic predictability affect the availability of these cues. The cues specifically tested for were coarticulatory effects of the target vowel on C_1, measured in the form of burst/frication duration and center of gravity (COG) of C_1.

Gender did not seem to have an effect on the acoustic results other than men having lower COG measurements for some consonants, which is expected given vocal tract length differences. However, male participants were shown to devoice more than the female participants, which confirms what Imai (2010) also found in younger speakers. What is interesting from the devoicing results, however, is where the observed difference between men and women came from. With tokens in devoicing environments having devoicing rates of essentially 100%, the difference in devoicing rates was clearly from the voicing tokens. An analysis of just the voicing tokens showed that devoicing rates were significantly different for high predictability environments but not low predictability environments. In other words, predictability also seems to affect devoicing rates, although only in men.

With respect to the issue of lengthening, duration measurements showed that lengthening is observable only in non-fricatives. Devoicing generally had no effect on fricatives with the exception of /s/, which shortened in devoiced contexts instead. This contrasts with Kondo (1997), who found lengthening effects of devoicing for all consonants. The observed difference is most likely because the current study compares C_1 duration in voicing versus devoicing environments (e.g., /kugi/ vs. /kuki/), whereas Kondo (1997) compares the duration of C_1 from voiced and devoiced instances of the same devoiceable environments (e.g., [kuʦuʃita] vs. [kuʦu̥ʃita]). Kondo was able to do this because the stimuli used contained consecutive devoicing environments, which may have led to different gestural timing patterns.

The fact that C_1 lengthening is dependent on the manner of the consonant suggests that it is not an obligatory process whose goal is to maintain mora-timing (Han, 1994). Furthermore, the fact that /ʧ/ lengthened while /ɸ, ç/ did not despite similar durations suggests that C_1 lengthening is not a recoverability-conditioned process, but rather is physiological in nature, where the lengthening observed in stops and affricates is due to the relatively high subglottal pressure compared to fricatives. Previous articulatory studies have found that in devoiced syllables with stop C_1, abduction peaks occur after the stop release (Weitzman et al., 1976) with distinct laryngeal muscular activities associated with the C_1 and the devoiced vowel (Simada et al., 1991, as cited in Kondo, 1997). In devoiced syllables with fricative or affricate C_1, however, the laryngeal activities are indistinguishable between the C_1 and devoiced vowel. Although the affricate [ʧ] was found to pattern with [k] in the current study, the results nevertheless suggest manner-conditioned differences in how high vowels become devoiced.

On the other hand, devoiced /s/ tokens showed significant shortening while /ʃ/ did not, despite similar durations of $\sim$100 ms. The reason for shortening in /s/ can be explained in terms of recoverability. Since the devoiced vowel after /s/ is highly predictable, the vowel can be deleted, and /s/ needs only to be long enough to signal the consonant's identity. As for why /ʃ/ cannot shorten, COG results must be discussed first, which are summarized below in Table VII.

TABLE VII: Summary of COG results.

		vowel (/ɯ/)	*devoicing*	*gender (male)*
ç	COG1	—	raising	n.s.
	COG2	—	raising	n.s.
ɸ	COG1	—	raising	n.s.
	COG2	—	raising	n.s.
s	COG1	—	n.s.	n.s.
	COG2	—	raising	n.s.
ʧ	COG1	—	n.s.	lowering
	COG2	—	raising	lowering
ʃ	COG1	lowering	n.s. (lowering for /ʃi/ in women)	lowering
	COG2	lowering	raising	lowering
k		lowering	raising	n.s.

C_1V coarticulation was predicted to lower the COG of C_1, while C_1C_2 coarticulation, where C_2 is alveolar, was predicted to raise the COG of C_1. Since the vowels in /çi , ɸu/ essentially have the same places of articulation as the consonants, C_1V coarticulation was expected to lower COG values for /ç, ɸ/ due to weakening constriction. Devoicing, however, had a raising effect for the two consonants for both COG1 and COG2. This suggests that vowel gestures were not maintained as in the case of voiced tokens from the very beginning. Because there is no intervening vocalic target, constrictions can be made tighter, leading to a rise in COG. Devoiced vowels, therefore, seem to be deleted in these contexts.

/s/ showed only that devoicing has a raising effect on COG2, suggesting coarticulation with the following C_2. Since devoicing had a raising effect on all C_1, the raising effect alone is not enough to distinguish between devoiced vowels being unphonated and deleted, but together with the shortening effect of devoicing on /s/, it seems likely that the vowel is deleted.

/ʧ/ results can be compared directly with /ʃi/ results, since the two consonants share a place of articulation and the vowel that follows. Although the effect was limited to female speakers, devoicing had a significant lowering effect on /ʃi/ tokens, but not on /ʧ/ tokens. If the lowering effect of devoicing on /ʃi/ is interpreted to mean increased coarticulation, where the palatal gesture of the vowel shifts the tongue back and enlarges the front oral cavity, then the lack of a comparable effect on /ʧ/ suggests that a similar effort is not being made to aid recoverability, at least in the case of female speakers. The acoustic results alone, however, are admittedly unclear, and perhaps an articulatory study would help clarify further whether the vowel is deleted or unphonated after /ʧ/.

/ʃ/ results showed both C_1V and C_1C_2 coarticulation. First, /u/ had a lowering effect on both COG1 and COG2, regardless of devoicing status, and although the effect was limited to female speakers, devoiced /ʃi/ tokens also showed a lowering effect. Tsuchida (1994), who analyzed speech recorded from three female speakers also reports a similar lowering effect of devoicing during the first half of /ʃi/. Tsuchida, however, also found devoicing to have a lowering effect on /ʃu/ throughout the entire C_1, which seemed to aid Japanese listeners in identifying the vowel in devoiced tokens even more successfully than in voiced tokens. This further lowering effect of devoicing on /ʃu/ tokens was not found in the current study. One possible explanation for the diverging results is that the analysis window used for COG measurements was longer in the current study (10 ms vs. 20 ms). It also seems likely that the differences are due to changes in the Japanese language itself, where younger speakers produce /u/ with more lip protrusion in general (Nogita et al., 2013), making further protrusion in devoiced /ʃu/ tokens more difficult or unnecessary.

Second, devoicing had a raising effect on COG2, suggesting C_1C_2 coarticulation. However, devoiced /ʃu/ tokens were still lower than devoiced /ʃi/ tokens. The persistent effect of /u/ suggests that there is an oral vowel gesture (lingual, labial, or both) that lengthens the front oral cavity. However, the raising effect of devoicing suggests that there is a lack of an intervening vocalic gesture that blocks C_1C_2 coarticulation. The two results can be reconciled if the lingual and labial vocalic gestures are thought of independently. Shaw and Kawahara (2018) investigated /u/ devoicing using electromagnetic articulography (EMA) and found that there is often no lingual gesture associated with devoiced vowels, and thus propose that the vowel must be deleting. However, the study did not investigate labial gestures, and as previously mentioned, /u/ is often rounded in young Japanese speakers (Nogita et al., 2013), which means that the labial gesture can be retained while the lingual gesture is lost. The COG results of /ʃ/ suggest that this is indeed what is happening. Devoiced vowels lose their lingual gestures, allowing /ʃ/ to coarticulate with C_2, but /u/ also retains its labial gesture, leading to lower COG values that help distinguish /u/ from /i/. The lowering effect of /u/ on /ʃ/ was also reported by Beckman and Shoji (1984) and Tsuchida (1994), and both studies also found that the coarticulatory effect aided identification of the vowel for Japanese listeners. The /ʃ/ results, therefore, suggest that devoiced vowels are neither simply unphonated nor completely deleted, but rather *reduced* in the sense that gestures associated with the vowel are lost incrementally. This retention of vocalic oral gestures also helps explain why /s/ shortened in duration, while /ʃ/ did not despite being similar in length. /s/ does not need to carry coarticulatory information of the following devoiced vowel because the vowel is predictable. /ʃ/, however, cannot shorten because the frication noise must be long enough to carry the coarticulatory cues of the devoiced vowel.

Lastly, the single COG measurement for /k/ showed that /u/ had a significant lowering effect, or perhaps more accurately that /i/ had a significant raising effect. The large spectral difference is most likely due to /k/-fronting that results from coarticulation with the following /i/, and positing the presence of coarticulatory effects even in devoiced tokens allows /k/ to be grouped with /ʃ/. However, the large COG difference of ∼2200 Hz between the burst noises of /ki/ and /ku/ is nearly three times the differences of ∼600–800 Hz observed for /ʃ/ in the current study and nearly six times the 400 Hz spectral difference reported in Beckman and Shoji (1984), differences to which Japanese speakers were shown to be sensitive. Given such a large spectral difference, it seems possible that velar fronting is categorical (i.e,. [k^j]) rather than a relative fronting (i.e., [k̟]) as was assumed throughout the current study. It is also possible then, that the spectral difference is not due to coarticulation with the vowels *per se*, but rather because the consonants preceding /i, u/ are simply different phonemes, namely /k^j, k/, respectively (an observation also made in Maekawa and Kikuchi (2005), as made evident by the transcription convention employed). If this is indeed the case, the devoiced vowels after [k^j, k] become highly predictable. A recalculation of entropy and surprisal for /k, k^j/ from the “Core” subset of Corpus of Spontaneous Japanese (Maekawa, 2003) showed that when only high vowels are considered, both entropy and surprisal are zero for /ku/ and near-zero for /kʲ/ (entropy = 0.036; surprisal = 0.005). While a high back vowel can follow /kʲ/, it is almost always the long vowel /uː/, which typically does not devoice. Even in the case of loanwords where /kʲ/ is followed by /u/, there is generally an alternative pronunciation as simply /kʲi/, showing again that a short high back vowel is dispreferred after /kʲ/ in the language (Shogakukan, 2013). It is admittedly difficult to tell apart based on the single acoustic measurement used in the current study whether the apparent fronting effect is due to C_1V coarticulation or simply due to different C_1, and perhaps an articulatory study looking at the oral gestures during closure would be helpful. Regardless of whether the /k/ results are perceptibility- or predictability-driven, however, both interpretations are compatible with the recoverability-based framework being proposed in this study.

V. CONCLUSION

The results of the current study provide further evidence that Japanese high vowel devoicing can result in complete deletion of the vowel (Pinto, 2015; Shaw and Kawahara, 2018), and the COG results in particular suggest that devoiced vowels are less likely to be deleted completely when they are unpredictable (i.e., after /ʃ/ and perhaps /k/), supporting the results of previous studies which showed that coarticulation between segments are controlled to aid perceptibility (Silverman, 1997; Chitoran et al., 2002). The results also provide novel insight into recoverability-driven coarticulation in that speakers not only retain the perceptibility of a devoiced vowel throughout the consonant when recoverability is in jeopardy (i.e., /ʃ/) but that they also do the opposite, where the vowel is deleted completely because it is highly predictable from the phonotactics (i.e., after /ç, ɸ/ in particular and possibly /s, ʧ/) and additional coarticulatory cues are unnecessary for recovery.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1524133. Thanks to Lisa Davidson, Shigeto Kawahara, Laura Koenig, and two anonymous reviewers for helpful comments and suggestions, as well as to audiences at ASA 167, LabPhon 14, NYU PEP Lab, and Keio University.

FOOTNOTES

[1] Rendaku is a morphophonological process in Japanese compounds, where the initial consonant of the second member of the compound becomes voiced (e.g., |ʦuki + ʦ̲uki| → /ʦukiʣ̲uki/ ‘month after month (moon + moon)’).

[2] See supplementary material at [URL will be inserted by AIP] for a full list of stimuli and carrier sentences.

[3] An analysis of consonant durations in the Corpus of Spontaneous Japanese revealed that there is no significant duration difference between [ʧ] and [ɸ] in voiced contexts (∼65 ms; $p = 0.891$), and between [ʧ] and [ç] in devoiced contexts (∼75 ms; $p = 0.475$).

[4] Although, see Kumagai (1999) whose EPG study found that palatal constriction is more fronted before [ɸ] in devoiced syllables.

REFERENCES

Baayen, R. Harald., Douglas J. Davidson, and Douglas. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59(4):390–412.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68(3):255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67(1):1–48.

Beckman, Mary. 1982. Segmental duration and the ‘mora’ in Japanese. *Phonetica* 39(2-3):113–135.

Beckman, Mary, and A. Shoji. 1984. Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica* 41(2):61–71.

Boersma, Paul. 2009. Cue constraints and their interactions in phonological perception and production. In *Phonology in perception*, ed. Paul Boersma and Silke Hamann, 55–110. Berlin: Mouton de Gruyter.

Browman, Catherine P., and Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4):155–180.

Chitoran, Ioana, Louis Goldstein, and Dani Byrd. 2002. Gestural overlap and recoverability: Articulatory evidence from Georgian. In *Papers in Laboratory Phonology VII*, ed. Natasha Warner and Carlos Gusshoven. Berlin: Mouton de Gruyter.

Cutler, Anne, Takashi Otake, and James M. McQueen. 2009. Vowel devoicing and the perception of spoken Japanese words. *J. Acoust. Soc. Am.* 125(3):1693–1703.

Faber, Alice, and Timothy J. Vance. 2000. More acoustic traces of “deleted” vowels in Japanese. In *Japanese/Korean Linguistics*, ed. Mineharu Nakayama and Charles J. Jr. Quinn, volume 9, 100–113.

Forrest, Karen, Gary Weismer, Paul Milenkovic, and Ronald N. Dougall. 1988. Statistical analysis of word-initial voiceless obstruents: preliminary results. *J. Acoust. Soc. Am.* 84(1):115–123.

Fowler, Carol A., and Elliot Saltzman. 1993. Coordination and coarticulation in speech production. *Lang. Speech* 36(2-3):171–195.

Fujimoto, Masako. 2015. Vowel devoicing. In *Handbook of Japanese Phonetics and Phonology*, ed. Haruo Kubozono, chapter 4. Mouton de Gruyter.

Fujimoto, Masako, Emi Murano, Seiji Niimi, and Shigeru Kiritani. 2002. Differences in glottal opening patterns between Tokyo and Osaka dialect speakers: Factors contributing to vowel devoicing. *Folia Phoniatrica et Logopedia* 54(3):133–143.

Hamann, Silke, and Anke Sennema. 2005. Acoustic differences between German and Dutch labiodentals. *ZAS Papers in Linguistics* 42:33–41.

Han, Mieko S. 1994. Acoustic manifestations of mora timing in Japanese. *J. Acoust. Soc. Am.* 96(1):73–82.

Hayes, Bruce. 1999. Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In *Functionalism and Formalism in Linguistics*, ed. Michael Darnell, Edith

Moravcsik, Michael Noonan, Frederick J. Newmeyer, and Kathleen M. Wheatley, 243–285. Amsterdam: John Benjamins.

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

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

Imai, Terumi. 2010. An emerging gender difference in Japanese vowel devoicing. In *A Reader in Sociolinguistics*, ed. Dennis Richard Preston and Nancy A. Niedzielski, volume 219, chapter 6, 177–187. Walter de Gruyter.

Ito, Junko. 1986. Syllable Theory in Prosodic Phonology. Doctoral Dissertation, University of Massachusetts, Amherst. Published 1988. Outstanding Dissertations in Linguistics series. New York: Garland.

Ito, Junko, and Armin Mester. 2003. Lexical and postlexical phonology in Optimality Theory: evidence from Japanese. *Linguistische Berichte* 11:183–207.

Ito, Junko, and Armin Mester. 2015. Sino-japanese phonology. In *Handbook of Japanese Phonetics and Phonology*, ed. Haruo Kubozono, chapter 7. Mouton de Gruyter.

Kindaichi, Haruhiko. 1995. *[Japanese Accent Dictionary]*. Sanseido.

Kiss, Zoltán, and Zsuzsanna Bárkányi. 2006. A phonetically-based approach to the phonology of /v/ in Hungarian. *Acta Linguistica Hungarica* 53(2-3):175–226.

Kondo, Mariko. 1997. Mechanisms of vowel devoicing in Japanese. Doctoral Dissertation, University of Edinburgh.

Kondo, Mariko. 2005. Syllable structure and its acoustic effects on vowels in devoicing. In *Voicing in Japanese*, ed. Harry van der Hulst, Jan Koster, and Henk van Riemsdijk, 229–246. Mouton de Gruyter.

Kubozono, Haruo. 2015. Loanword phonology. In *Handbook of Japanese Phonetics and Phonology*, ed. Haruo Kubozono, chapter 8, 313–362. Mouton de Gruyter.

Kumagai, Shuri. 1999. Patterns of linguopalatal contact during Japanese vowel devoicing. In *Int. Cong. Phon. Sci. 14*. 375–378.

Kurisu, Kazutaka. 2001. The Phonology of Morpheme Realization. Doctoral Dissertation, University of California, Santa Cruz.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmertest: Tests in linear mixed effects models. *J. Stat. Software* 82(13):1–26.

Maekawa, Kikuo. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. *Proceedings of the ISCA & IEEE workshop on spontaneous speech processing and recognition (SSPR)* .

Maekawa, Kikuo, and Hideaki Kikuchi. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In *Voicing in Japanese*, ed. Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara. Mouton de Gruyter.

Mattingly, Ignatius G. 1981. Phonetic representation and speech synthesis by rule. In *The cognitive representation of speech*, ed. J. Myers, J. Laver, and Anderson J., 415–420. North-Holland Publishing Company.

McCarthy, John J. 1999. Sympathy and phonological opacity. *Phonology* 16(3):331–399.

Nielsen, Kuniko Y. 2008. Word-level and feature-level effects in phonetic imitation. Doctoral Dissertation, University of California, Los Angeles.

Nittrouer, Susan., Michael Studdert-Kennedy, and Richard S. McGowan. 1989. The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *J. Speech Hear. Res.* 32(1):120–132.

Nogita, Akitsugu, Noriko Yamane, and Sonya Bird. 2013. The Japanese unrounded back vowel /ɯ/ is in fact rounded central/front [ʉ - ʏ]. *Ultrafest VI*.

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

Ogasawara, Naomi, and Natasha Warner. 2009. Processing missing vowels: Allophonic processing in Japanese. *Lang. Cog. Processes* 24(3):376–411.

Okamoto, Shigeko. 1995. “Tasteless” Japanese: Less “feminine” speech among young Japanese women. In *Gender articulated: Language and the socially constructed self*, ed. Kira Hall and Mary Bucholtz, 297–325. New York: Routledge.

Pinto, Francesca. 2015. High vowels devoicing and elision in japanese: a diachronic approach. In *Int. Cong. Phon. Sci. 18*.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27:379–423.

Sharoff, Serge. 2008. Lemmas from the internet corpus. URL `http://corpus.leeds.ac.uk/frqc/internet-jp.num`.

Shaw, Jason, and Shigeto Kawahara. 2017. Effects of surprisal and entropy on vowel duration in Japanese. *Lang. Speech* 0(0):0023830917737331.

Shaw, Jason, and Shigeto Kawahara. 2018. The lingual articulation of devoiced /u/ in Tokyo Japanese. *J. Phon.* 66:100–119.

Shogakukan. 2013. Daijisen Zoubo/Shinsouban (Digital Version). URL `http://dictionary.goo.ne.jp/`.

Silverman, Daniel. 1997. Phasing and Recoverability. Doctoral Dissertation, University of California, Los Angeles.

Simada, Zyun'ici B., Satoshi Horiguchi, Seiji Niimi, and Hajime Hirose. 1991. Devoicing of Japanese /u/: An electromyographic study. In *Int. Cong. Phon. Sci. 12*.

Tamaoka, Katsuo and Shogo Makioka. 2004. Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers* 36(3):531–547.

Tateishi, Koichi. 1989. Phonology of Sino-Japanese morphemes. In *University of Massachusetts occasional papers in linguistics 13*, 209–235. Amherst: GLSA Publications.

Tsuchida, Ayako. 1994. Fricative-vowel coarticulation in Japanese devoiced syllables: Acoustic and perceptual evidence. *Working Papers of the Cornell Phonetics Laboratory* 9:183–222.

Tsuchida, Ayako. 1997. Phonetics and phonology of Japanese vowel devoicing. Doctoral Dissertation, Cornell University.

Tsuchida, Ayako, Shigeru Kiritani, and Seiji Niimi. 1997. Two types of vowel devoicing in Japanese: Evidence from articulatory data. *J. Acoust. Soc. Am.* 101(5):3177.

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

Varden, J. Kevin. 1998. On high vowel devoicing in standard modern Japanese. Doctoral Dissertation, University of Washington.

Varden, J. Kevin. 2010. Acoustic correlates of devoiced Japanese vowels: Velar context. *J. Eng. Am. Lit. Ling.* 125:35–49.

Varden, J. Kevin, and Tsutomu Sato. 1996. Devoicing of Japanese vowels by Taiwanese learners of Japanese. *Proceedings of Int. Conf. on Spoken Lang. Processing* 96(2):618–621.

Warner, Natasha, and Takayuki Arai. 2001a. Japanese mora-timing: A review. *Phonetica* 58(1-2):1–25.

Warner, Natasha, and Takayuki Arai. 2001b. The role of the mora in the timing of spontaneous Japanese speech. *J. Acoust. Soc. Am.* 109(3):1144–1156.

Weitzman, Raymond S., Masayuki Sawashima, Hajime Hirose, and Tatsujiro Ushijima. 1976. Devoiced and whispered vowels in Japanese. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics* 10:61–79.

Yoshioka, Hirohide. 1981. Laryngeal adjustment in the production of the fricative consonants and devoiced vowels in Japanese. *Phonetica* 38(4):236–351.

Yoshioka, Hirohide, Anders Löfqvist, and Hajime Hirose. 1980. Laryngeal adjustments in Japanese voiceless sound production. *J. Acoust. Soc. Am.* 67(S1):S52-S53.

Zsiga, Elizabeth. 2000. Phonetic alignment constraints: consonant overlap in English and Russian. *J. Phon.* 28(1):69–102.