Effects of phonotactic predictability on sensitivity to phonetic detail

James Whang
MARCS Institute for Brain, Behaviour & Development
ARC Centre of Excellence for the Dynamics of Language
research@jameswhang.net

Abstract

Japanese speakers systematically devoice or delete high vowels [i, u] between two voiceless consonants. Japanese listeners also report perceiving the same high vowels between consonant clusters even in the absence of a vocalic segment. Although perceptual vowel epenthesis has been described primarily as a phonotactic repair strategy, where a phonetically minimal vowel is epenthesized by default, few studies have investigated how the predictability of a vowel in a given context affects the choice of epenthetic vowel. The present study uses a forced-choice labeling task to test how sensitive Japanese listeners are to coarticulatory cues of high vowels [i, u] and non-high vowel [a] in devoicing and non-devoicing contexts. Devoicing contexts were further divided into high-predictability contexts, where the phonotactic distribution strongly favors one of the high vowels, and low-predictability contexts, where both high vowels are allowed, to specifically test for the effects of predictability. Results reveal a strong tendency towards [u] epenthesis as previous studies have found, but the results also reveal a sensitivity to coarticulatory cues that override the default [u] epenthesis, particularly when the vowel is in a context that is less predictable. Previous studies have shown that predictability affects phonetic implementation during production, and this study provides evidence that listeners are also similarly affected by predictability.

Key words: perceptual repair, phonotactics, predictability, Japanese

1. Introduction

The current study investigates Japanese listeners and the role of phonotactic predictability in how illicit consonant clusters are repaired. While it is commonly thought that Japanese listeners use [u] epenthesis by default because it is the shortest (and thus phonetically minimal) vowel (Dupoux et al., 1999, 2011), the current study proposes that the choice of epenthetic vowel in Japanese listeners rely on a combination of phonotactic predictability and attention to phonetic cues, based on experience with recovering high vowels that are systematically devoiced or deleted (Shaw & Kawahara, 2018; Whang, 2018) in their language. To distinguish the respective roles of phonotactic prediction and phonetic cue perception, participants are presented with conflicting phonotactic and phonetic information, allowing insight into which of the two they prioritize.

1.1. Effects of predictability during production

Phonetic cues are often weakened for segments that are predictable from a given context. Exemplar-based approaches to phonology (Bybee, 2006; Ernestus, 2011; Pierrehumbert, 2001) have long noted that it is often the most frequent lexical items that are targeted for devoicing due to their predictability. Building on this line of research, Hall et al. (in preparation) argue that phonological systems tend to reduce segments in predictable contexts because enhancing the cues would require additional effort while contributing little to successful lexical access (by the listener). For example, word-final coda contrasts are often neutralized cross-linguistically because segments become more predictable as the listener processes more and more of the target item during lexical access. This means that word-final codas contribute less to identifying the target lexical item. Rather than enhancing the weak cues of an already predictable segment, phonological systems choose to enhance cues of segments in unpredictable positions instead, such as in the case of word-initial obstruent aspiration in English (e.g., /$\underline{\text{p}}$ik/ → [$\underline{\text{p}}^{\text{h}}$ik] vs. /s$\underline{\text{p}}$ik/ → [s$\underline{\text{p}}$ik]). Predictability, specifically phonotactic predictability has also been shown to have similar effects on high vowel devoicing in Japanese. Shaw and Kawahara (2018) found that devoiced high vowels in Japanese often deleted completely, leading to consonant clusters, and Whang (2018) found that predictability had a noticeable effect on the likelihood of deletion, where highly predictable vowels deleted while less predictable vowels retained oral gestures, providing coarticulatory cues.

1.2. Effects of phonotactic knowledge on speech perception

If it is the case that speakers are varying the amount of phonetic cues depending on the target segment's predictability in a given context, the question that naturally follows is whether listeners similarly vary their attention to phonetic cues based on predictability. Numerous studies have shown that expectations that stem from language experience affect how listeners utilize phonetic cues. For example, listeners are often insensitive to phonetic cues that are not contrastive in their native language. French listeners have difficulty contrasting short versus long vowels (Dupoux et al., 1999), English listeners have difficulty perceiving tonal contrasts (So & Best, 2010), Japanese listeners have difficulty contrasting /l/ versus /r/ because neither are phonemes of the language (Flege et al., 1996), and so on. Conversely, listeners are also attuned to cues that are useful in their native language. For example, Korean listeners are more sensitive to V-to-C formant transition cues than English listeners (Hume et al., 1999) because coda obstruents are obligatorily unreleased in Korean while they are optionally released in English (Kang, 2003), making the transitional cue more useful to Korean listeners for recovery of the coda consonant than to English listeners, who have the option of waiting for the release of the coda obstruent.

Language experience also shapes the perception of phonetic cues. Pitt and McQueen (1998) showed that listeners are biased towards identifying phonetically ambiguous segments as segments with higher phonotactic probability (i.e., phonotactically more predictable). Phonotactic knowledge also seems to play an important role in other domains as well. When processing nonce words, sequences with higher phonotactic probabilities are processed faster (Vitevitch et al., 1997; Vitevitch & Luce, 1999). On the other hand, lexical items with high phonotactic probabilities are processed slower than lexical items with low phonotactic probabilities, presumably because high phonotactic probability in lexical items means that there are also that many more similar lexical items, utimately slowing down lexical access (Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994; Vitevitch & Luce, 1998). The question then is, which process takes precedence? The answer to this question seems to depend on the task. In general, listeners seem to prioritize the use of lexical knowledge, relying on their phonotactic knowledge only when lexical activation fails (Shademan, 2006; Vitevitch & Luce, 1999; although see Myers (2015) and related works showing significant contributions of neighborhood density on wordlikeness judgments in non-English languages). Additionally, Mattys et al. (2005) investigated whether participants pay more attention to lexical and sublexical (segmental and prosodic) segmentation cues when they are in conflict. The results showed again that lexical cues are prioritized and that listeners rely on sublexical cues when lexical context or information cannot be accessed due to noise or absence. The current study adds to this line of work on "top-down" effects by investigating the interaction between two kinds of sublexical information, namely phonotactic predictability and fine-grained phonetic cues, using high vowel epenthesis in Japanese as a test case.

1.3. Perceptual repair by Japanese listeners

In the now well-known study commonly referred to as the "*ebzo* test" (Dupoux et al., 1999), French and Japanese speakers were presented with acoustic stimuli with the high back rounded vowel [u] of varying durations ranging from 0 ms to 90 ms occurring between two consonants (e.g., [ebzo] → [ebuːzo]). The stimuli were designed so that when there is no vowel in the stimuli, the result is a sequence that is phonotactically legal in French but illegal in Japanese. Their results showed that while French speakers could accurately distinguish the vowel-less from vowel-ful tokens, Japanese speakers were essentially "deaf" to such differences, erring heavily towards misperceiving what the authors call an "illusory" vowel. On the other hand, French speakers were unable to accurately perceive vowel length, with which the Japanese participants had little trouble perceiving. The authors propose that the results are due to phonotactic differences in French and Japanese, where Japanese listeners perceive a non-existent vowel between two consonants because Japanese phonotactics disallows heterorganic consonant clusters. French listeners, on the other hand, were insensitive to vowel length because it is not contrastive in French. The authors further argue that there is a "top-down" phonotactic effect on perception, where phonotactically illegal sequences are automatically perceived as the nearest legal sequence rather than repaired at a higher, abstract phonological level.

Dehaene-Lambertz et al. (2000) also tested the illusory vowel epenthesis effect in an event-related potential (ERP) study. In this study, Dehaene-Lambertz et al. carried out experiments similar to that of Näätänen et al. (1997), where electrophysiological responses have been shown to be sensitive to phoneme categories. Dehaene-Lambertz et al. looked at how mismatch negativity (MMN) responses in Japanese and French speakers differ in the absence versus presence of a vowel in the same kind of sequences as those in Dupoux et al. (1999). The experiments followed an oddball paradigm where in one trial a sequence that is legal in both languages was presented as the standard (e.g., [igumo]) and one that is illegal only in Japanese as the deviant (e.g., [igmo]). The reverse was presented in a separate trial. Although the results reported collapsed the trials, the ERP results generally showed that Japanese speakers are insensitive to the differences between the vowel-ful and vowel-less items, while French speakers are, supporting the behavioral results from the original study by Dupoux et al. (1999). A related fMRI study by Jacquemot et al. (2003), also found similar but slightly weaker results. Jacquemot et al. report that in an AAX task (A-stimulus presented twice before X-stimulus), neural activity increased whenever the X stimulus was different from the A stimulus for both Japanese and French participants. This was true regardless of whether or not the acoustic difference was phonologically contrastive in the language, although neural activation was significantly greater when the acoustic contrasts were also phonologically contrastive.

A more recent study by (Dupoux et al., 2011) aimed to further bolster the automatic perceptual repair idea by also investigating European Portuguese, Brazilian Portuguese, and Japanese listeners. The reason for choosing the two dialects of Portuguese was that European Portuguese allows the same types of clusters as French, but Brazilian Portuguese has a strict CVCV phonotactic structure, leading to the expectation that their perception would be similar to that of Japanese listeners. The crucial difference between Brazilian Portuguese and Japanese is that in the former, the default epenthetic vowel is reported to be /i/ as opposed to the Japanese /u/. Since the quality of the epenthetic vowels are different in the two epenthesizing languages, the experiments were modified slightly from the 1999 study to enable identification of the perceived illusory vowels in the results. Like French listeners, the results showed that European Portuguese listeners did not have trouble distinguishing vowel-less from vowel-ful tokens. Japanese listeners, again, showed a tendency towards mistakenly recovering /u/ between consonant clusters. The results, however, additionally showed that Japanese listeners were also sensitive to [i]-coarticulation in the first consonant (i.e., $e\underline{b^j}zo$), recovering /i/ rather than /u/. By comparison, Brazilian Portuguese listeners tended to perceptually recover /i/ between illegal consonant clusters by default as expected, but did not show the same degree of sensitivity to [u]-coarticulation. Although the reasons for the disparity in sensitivity to coarticulatory cues were not discussed, the difference is likely due to Brazilian Portuguese listeners having little experience with a systematic high vowel devoicing process, leading them to underutilize coarticulatory cues relative to Japanese listeners.

1.4. Problems and solutions

The series of studies discussed above collectively suggest that there is a top-down imposition of the listeners native phonotactic grammar during perception. The experiments, however, would benefit from two particular refinements when considering Japanese listeners: using stimuli that are less foreign to Japanese listeners and controlling for the effects of high vowel devoicing in how Japanese listeners perceive certain consonant clusters.

First, the waveform and spectrogram examples of the stimuli used in the studies by Dupoux and colleagues reveal that the burst of C_1 (e.g., [b] in [ebzo, ebuzo]) were rather long, potentially biasing the participants to perceive a vowel. For example, Dupoux et al. (2011) shows that in a sequence like [agno], the voiced stop had a burst of at least 50 ms and contained formant-like structures. Japanese voiced stops, however, typically have burst durations of less than 20 ms (Kong et al., 2012). In addition, Japanese high vowels are inherently short, with an average duration of approximately 40 ms, but they can be as short as 20 ms (Han, 1994; Beckman, 1982). Taking the short burst and vowel durations of Japanese together, an atypically long burst with formant structures can be interpreted as containing a vowel, possibly confounding the independent effects of acoustic cues and phonotactic violations (Wilson et al., 2014). Furthermore, the closure duration of the voiced stop is also nearly 100 ms, which is closer to the geminate range than the singleton range in Japanese (Kawahara, 2006). Geminates are not known to affect high vowel devoicing in C_1 position, but geminate consonants in C_2 position have been shown to increase the likelihood of preceding vowels being phonated in Japanese regardless of whether the consonants are voiced (Maekawa & Kikuchi, 2005; Fujimoto, 2015). This means that stimuli with geminate-like obstruents in both C_1 and C_2 positions (e.g., [igba]) could have further biased Japanese participants toward expecting a vowel in the target context. While this is also a tendency that is phonotactically driven, it is unclear whether the primary driving force behind perceptual epenthesis in the experiments is the heterorganic clusters, the phonetic cues of geminate-like segments, or a combination of both.

Second, the stimuli used in the *ebzo* tests included a mix of environments in which high vowel devoicing is expected to occur in Japanese as well as non-devoicing environments. The results reported in these studies, however, make no distinction between the two types of environments. Japanese high vowel devoicing is a highly productive process that applies at rates above 80% in most contexts (Maekawa & Kikuchi, 2005), where high vowels lose at the least their phonation and at most delete completely when between voiceless obstruents (e.g., /masutaa/ → [mastaa] 'master'; Shaw & Kawahara, 2018). Given the life-long experience Japanese listeners have in recovering the often-deleted high vowels between consonant clusters, it is very likely that this phonological process has an effect that is independent of phonotactic violations in creating an expectation for a vowel, and the most straightforward remedy to this issue is to test and analyze devoicing and non-devoicing environments separately (e.g., [ezpo] vs. [espo]).

Furthermore, related to the division of devoicing and non-devoicing stimuli, the devoicing stimuli can be divided into low- and high-predictability sub-groups. Varden (2010) states what seems to be a prevalent assumption in the literature on Japanese high vowels, which is that since high vowels trigger allophonic variation for /t, s, h/ in the language (i.e., /t/ → [ʧi, ʦu]; /s/ → [ʃi, su]; /h/ → [çi, ɸu]), the high vowels need not be acoustically present in these contexts because they can be predicted with certainty from the allophonic consonant alone. Conversely, this also means that in environments where allophonic variation is not triggered (i.e., /p, k, ʃ/ → [pi, pu, ki, ku, ʃi, ʃu]), Japanese listeners would be more inclined to pay closer attention to the phonetic cues because they cannot predict the vowel with certainty. This effect of predictability on the perception of high vowels has long been assumed but never tested systematically. The current study, therefore, presents a perception experiment that specifically controls phonotactic predictability and investigates how it affects Japanese listeners' utilization of coarticulatory cues Japanese.

It should be noted that although [ʧ, ʦ], [ʃ, s], and [ç, ɸ] are traditionally analyzed as allophones of /t, s, h/ before /i, u/, respectively as discussed above, the current study regards them as phonemes with extremely skewed phonotactic distributions to more accurately reflect how the sounds are used in Japanese today. For example, minimal loan pairs such as [tiaː] 'tier' and [ʧiaː] 'cheer' suggest that words like 'cheer' contain an underlying /ʧ/ that surfaces faithfully, rather than an underlying /t/ that undergoes allophony. In fact, an analysis of the Corpus of Spontaneous Japanese (Maekawa & Kikuchi, 2005, see Table 3 below for more details) revealed that with the exception of /ʦ/, which still only precedes /u/, all other "allophones" (/ʧ, ɸ, s, ʃ/) can now precede all vowels or most vowels (/*çe/).

2. Materials and methods

The stimuli for this study are in the form $V_1C_1(V_T)C_2V_2$, where V_T is the target vowel and C_1 and C_2 are determined based on the stimulus group the token belongs to. The stimuli were divided into three groups: non-devoicing (NoDevoice) where vowel devoicing is not expected, low predictability (LoPredict) where both high vowels can occur and devoice, making coarticulatory cues necessary for recovery of a devoiced vowel, and high predictability (HiPredict) where phonotactic predictability is sufficient for recovery of a devoiced vowel, making coarticulatory cues less important. Below in Table 1 are the stimuli. Note that although it is more accurate to use the IPA symbols for the voiceless alveopalatal fricative [ɕ] and affricate [ʨ] in Japanese, the palatoalveolar symbols [ʃ, ʧ] are used throughout the current study to make [ʃ] more visually distinct from the palatal fricative [ç] and to make [ʧ] consistent in place with [ʃ].

Table 1: Stimuli for Experiment 2.

NoDevoice	eb_ko	ez_po	eg_to	ob_ke	oz_pe	og_te
LoPredict	ep_ko	eʃ_po	ek_to	op_ke	oʃ_pe	ok_te
HiPredict	eɸ_ko	es_po	eç_to	oɸ_ke	os_pe	oç_te

There were 252 stimulus items in total. The stimulus forms shown in Table 1 were first recorded by a trained, non-Japanese-speaking, English-Hungarian bilingual phonetician in a sound-attenuated booth with stress on the initial vowel and with /i, u, a/ as target vowels (V_T). /a/ was included as a target vowel because it is a low vowel that typically does not devoice in Japanese, and also to test whether Japanese listeners are sensitive to coarticulatory cues of all vowels or just high vowels. Attempts were made to record the stimuli with two native Japanese speakers, but both speakers had difficulties keeping high vowels voiced in devoicing contexts, and even when they were successful in producing voiced high vowels in devoicing contexts, either the burst durations were too short to manipulate or the target vowel was stressed.

For each recording, the target vowels were manipulated by inserting or removing whole periods to achieve a duration of ∼40 ± 5 ms. From each of the recordings, four additional tokens were created by removing from right to left, half of V_T (splice-1), the remaining half of V_T (splice-2), half of the C_1 burst/frication noise (splice-3), then the remaining half of the C_1 burst/frication noise leaving only the closure for stops and ∼15 ms for fricatives (splice-4). An example of how the splicing was done is shown in Figure 1 below with the token [ekuto].

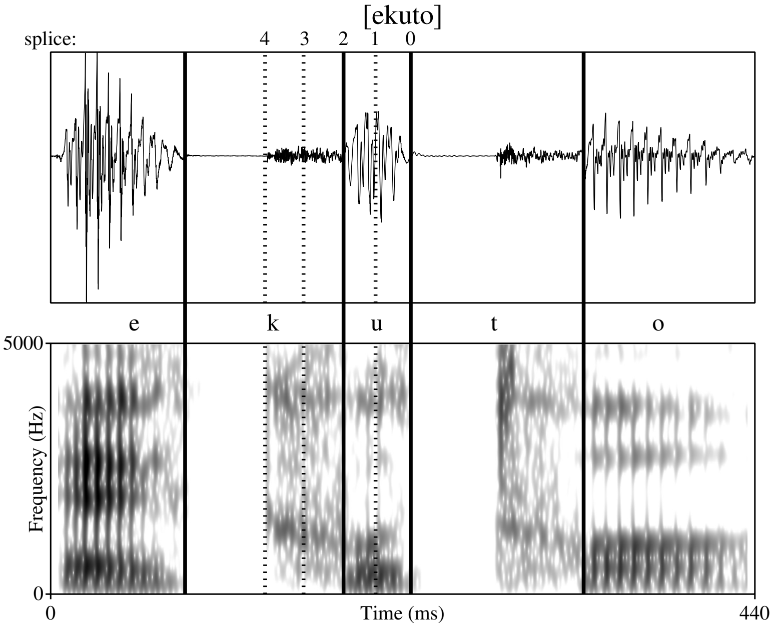

Figure 1: Example of token splicing: [ekuto].

The result of the splicing process is a gradual decrease of vowel coarticulatory information available in the burst/frication noise of C_1. Stop bursts in particular were manipulated to test whether it is phonotactic predictability or interpretation of phonetic information that drive illusory vowel epenthesis, since sensitivity to and interpretation of stop bursts as signaling the presence of a vowel is reported not just in Japanese (Furukawa, 2009; Whang, 2016) but in Korean (Kang, 2003) and English (Davidson & Shaw, 2012; Hsieh, 2013) as well.

Naturally produced, vowel-less tokens were also recorded for each stimulus form to test how it differs in perception from the spliced vowel-less stimuli (splice-2), which have traces of coarticulation from the target vowel on the surrounding consonants.

2.1. Participants

Twenty-nine monolingual Japanese listeners (16 women, 13 men) were recruited for the perception experiment 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 or had been overseas within a year prior to the experiment. All participants were compensated for their time.

2.2. Procedure

The experiment follows the forced-choice vowel labeling task from Dupoux et al. (2011). The participants were told that they would be listening to foreign words over headphones and that they would have 5 seconds to choose one of four answer choices that best matches the word they heard. The stimuli were presented through noise-isolating headphones, and answer choices that give the vowel-less and various vowel-ful spellings of the stimulus that just played were presented on screen simultaneously (e.g., [epuko] → <epko>, <epako>, <epiko>, <epuko>, where the <> notation indicates orthographic representation). Participants selected their answer choices by using arrow keys on a keyboard (i.e., ↑ ↓ ← →). A typical answer-choice screen is shown below in Figure 2.

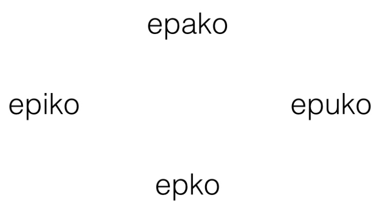

Figure 2: Answer choice screen for [epVko], where V = /a, i, u, ∅/.

While it is true that Japanese orthography is a syllabic system, most Japanese speakers are quite comfortable with the Latin alphabet, not only because of frequent exposure to loanwords but also because of the keyboards used for word processing. There are currently two main input methods—direct input (one key = one syllabic character) and conversion (QWERTY keyboard used to input CV combinations which are then converted to the corresponding syllabic character)—and the conversion method is commonly more preferred, and thus participants are expected to be comfortable with answer choices presented in the Latin alphabet. The experiment was designed to continue as soon as the participant made an answer choice.

2.3. Defining predictability

Before discussing its effects, predictability should be defined. Predictability can be quantified using two Information-Theoretic (Shannon, 1948) measures: *surprisal* and *entropy*. Both measures are calculated based on the conditional probabilities of vowels after a given consonant (i.e., $\Pr(v \mid C_1\_)$). Surprisal is the negative $\log_2$ probability ($-\log_2 \Pr(v \mid C_1\_)$) and indicates the amount of information (effort) necessary to predict a vowel after a given C_1. Entropy is the weighted average of surprisal in a given context ($\sum \Pr(v \mid C_1\_) * (-\log_2 \Pr(v \mid C_1\_))$) and indicates the overall level of uncertainty in a given context due to competition amongst other possible vowels. Due to expectations stemming from experience with high vowel devoicing, when given a voiceless $\mathrm{C}_1\mathrm{C}_2$ sequence with no apparent intervening vowel, experience with high vowel devoicing informs the Japanese listener that there is most likely an underlying /i, u/ that was devoiced. 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 2 are entropy and surprisal measures calculated from all underlying biphone tokens in the "Core" subset of the Corpus of Spontaneous Japanese (Maekawa, 2003; Maekawa & Kikuchi, 2005) for the consonants used for devoicing stimuli in the current study. Entropy and surprisal never reach zero, showing that all consonant allow both /i, u/ to follow.

Table 2: Overall entropy for devoicing C_1 and surprisal of /i, u/.

	IPA	*Entropy*	*Surprisal /i/*	*Surprisal /u/*
low predictability	p	0.830	1.931	0.439
	k	0.980	1.264	0.777
	ʃ	0.221	0.052	4.819
high predictability	ɸ	0.095	6.362	0.018
	s	0.021	8.991	0.003
	ç	0.012	0.001	9.914

Given the two measures, listeners are more likely to predict vowels with low surprisal overall, but the listener is also more likely to consider other vowels in high entropy (uncertainty) environments. In other words, phonetic cues for high vowels can be used to counteract uncertainty. To give a concrete example, given a "high predictability" context such as after /s/, the listener can predict the vowel to be /u/ with near-zero effort (surprisal = 0.003) and have near-absolute certainty about the prediction. In "low predictability" contexts such as after /p/, although /u/ is the most likely vowel, the high level of uncertainty would lead the listener to listen for phonetic cues that either support or contradict the context-based prediction.

If C_1 is voiced in the given C_1C_2 sequence, it no longer constitutes a devoicing environment, and thus the Japanese listener must consider all five vowels of Japanese. Given a five vowel system, the maximum entropy is 2.322 ($-\log_2 p(0.2)$), and entropy measures for /b, g, z/ were 2.084, 1.570, and 1.889, respectively. To provide a thorough overview of how likely each of the five vowels is to follow the consonants used in the current study, presented below in Table 3 are the observed/expected ratios for all pertinent C_1V biphones, calculated from the underlying biphone tokens provided in the Corpus of Spontaneous Japanese (i.e., devoicing status ignored). Observed/Expected ratios (O/E) quantify how overrepresented or underrepresented a given biphone is by dividing the biphone's observed number of occurrences (O) by the biphone's expected number of occurrences (E) if consonants combined at random. The resulting O/E value indicates the magnitude of difference from the expected value. For example, O/E of 2 indicates that the biphone occurred twice as often as expected, while O/E of 0.5 indicates that the biphone occurred half as often as expected.

Table 3: Observed/expected (O/E) ratio of C_1V from CSJ. Highest O/E in bold.

	NoDevoice			LoPredict			HiPredict		
	b_	g_	z_	p_	k_	ʃ_	ɸ_	s_	ç_
_a	1.63	**3.44**	0.93	1.78	1.80	0.27	0.11	0.92	0.43
_i	0.79	0.31	0.00	0.65	1.12	**6.28**	0.10	0.04	**6.28**
_u	**4.14**	0.78	**4.67**	**2.86**	**2.24**	0.33	**9.01**	**5.42**	0.002
_e	1.24	0.75	2.30	0.49	0.97	0.003	0.12	0.90	0.006
_o	0.75	1.33	0.99	0.43	1.33	0.42	0.07	1.16	0.01

The O/E values show that /u/ is highly overrepresented in Japanese after most consonants. The exceptions are /g/ after which /a/ is the most common vowel, and /ʃ, ç/ after which /i/ is the most common vowel, which was also shown by the low entropy of /i/ after these consonants in Table 2.

2.4. Analysis and predictions

All statistical analyses were performed by fitting linear mixed effects models using the *lme4* package (Bates et al., 2015) for R (R Core Team, 2016). The statistical analyses assess vowel detection and vowel identification. Detection refers to how often participants report perceiving any vowel at all both in the presence and absence of vocalic segments. Identification refers to whether the vowel the participants perceive is in agreement with the acoustic vocalic information contained in the stimuli.

In the case of detection, accuracy is expected to be higher for the *NoDevoice* group (e.g., [ez_po]) and lower in the *LoPredict* and *HiPredict* groups (e.g., [eʃ_po] and [es_po], respectively). Since the phonological process of high vowel devoicing is nearly obligatory (Vance, 1987), it could bias Japanese speakers to expect a high vowel to be present between two voiceless consonants even when it is acoustically absent. Since the current experiment uses nonce-words, there is no underlying or lexical form to access. Devoiced and voiced sequences involving two voiceless obstruents in Japanese would map to the same phonotactically legal underlying form (e.g., [esupo] ≡ [esu̥po] ≡ [espo] → /esupo/). The devoiced and voiced sequences would all be regarded as legal, and the actual presence or absence of the vowel in the signal is readily ignored. While devoicing is possible in the *NoDevoice* environments, it is extremely rare (Maekawa & Kikuchi, 2005). Since only the vowel-ful token is legal in the language in non-devoicing contexts, a devoiced or vowel-less counterpart is not in an equivalence relationship (e.g., [sude] ≢ *[su̥de] ≢ *[sde] 'barehand'). Thus Japanese listeners are expected to be more sensitive to the presence versus absence of a medial vowel. Furthermore, regardless of the stimulus group, higher accuracy is expected in recognizing that there is no vowel as the burst/frication noise gets shorter, especially when there is no burst present.

In the case of identification, high accuracy is expected for the *LoPredict* group and lower accuracy for the *HiPredict* group. Japanese speakers have been shown to be sensitive to high vowel coarticulation in /ʃ/ (Beckman & Shoji, 1984), but this sensitivity is only useful when the vowel is unpredictable after a given C_1 (i.e., *LoPredict* group). Japanese listeners, therefore, should be sensitive at least to [i, u]-coarticulation in the *LoPredict* group but biased towards a single high vowel that most frequently follows C_1 *HiPredict* group regardless of coarticulation. Since there are four answer choices <i, u, a, ∅>, identification rates are expected to be at least 50% in the *LoPredict* group and approximately 25% in the *HiPredict* group. Furthermore, since /a/ rarely devoices in Japanese, <a> responses should be relatively low even for [a]-coarticulated tokens, defaulting instead to the most phonotactically probably vowel. The *NoDevoice* group is expected to show some effects of coarticulation, as was the case in Dupoux et al. (2011), but like the *HiPredict* group, /a/ should show little effect.

These predictions contrast with the account given by Dupoux and colleagues. According to Dupoux and colleagues, there are two mechanisms at play during illusory vowel epenthesis. First, perceptual repair is a one-step process where phonotactically illegal sequences are perceived as their repaired counterparts rather than being perceived accurately first then repaired to their phonotactically legal counterparts. What this means is that listeners do not have access to the source language's underlying form, making heterorganic C_1C_2 sequences and their repaired C_1VC_2 sequences equivalent for Japanese listeners. If this is correct, the prediction in terms of detection is that the rate of vowel detection between C_1C_2 and C_1VC_2 sequences should be statistically the same since the two sequences are equivalent.

Second, although Dupoux and colleagues argue that perceptual repair is triggered by phonotactic violations, the authors propose a phonetic repair strategy, where Japanese listeners default to epenthesizing /u/ because it is the shortest vowel in the language, whereas Brazilian Portuguese listeners epenthesize /i/ instead for the same reason (Dupoux et al., 2011). If the choice of the epenthetic segment is indeed based on the magnitude of phonetic change rather than phonotactic probability, no observable effect of phonotactic predictability is expected, since phonotactic knowledge merely flags repair sites but is not involved in the repair itself. If the repair strategy argued by Dupoux and colleagues is correct, vowel identification rates are predicted to suffer across all contexts whenever the coarticulated vowel is not /u/.

3. Results

Shown in Figure 3 below are the overall results of the experiment. Figure 3.A shows results for all C_1 and Figure 3.B for stop C_1 only, which consequently also results in the exclusion of all high-predictability tokens, since /ɸ, s, ç/ are all fricatives. The colors indicate the target vowels, and the solid and dashed lines indicate vowel detection and successful vowel identification rates, respectively. Vowel detection rates simply collapse all non-zero responses, whereas vowel identification rates only include cases where participant responses matched the coarticulated vowels in the stimuli (e.g., respond <epuko> for [epu̥ko]). The smaller the distance between two lines of the same color, the higher the proportion of successful vowel identification.

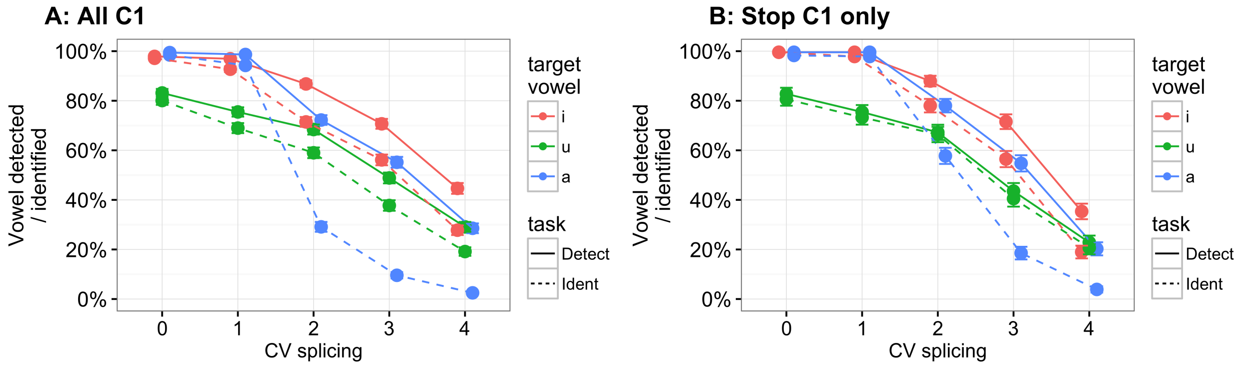

Figure 3: Vowel detection and identification rates with error bars by degree of splicing. CV splicing: 0 = full-CV, 1 = full-C half-V, 2 = full-C zero-V, 3 = half-C zero-V, 4 = zero-CV.

Figures 3.A and 3.B are qualitatively similar, where detection and identification rates fall as more of the C_1V_T information is spliced, and the most noticeable effect of including fricatives in 3.A is that identification rates are driven lower. In both figures, there are three things that stand out. First, detection rates for /u/ never reach 100% even when there is a full vowel of 40 ms present in the stimuli, suggesting that there is confusion between the presence and absence of /u/. Second, vowel detection rates never quite reach 0%, remaining above 20% even in the absence of any C_1 burst noise (Figure 3.B, splice-4), suggesting an overall confusion between vowel-fulness and vowel-lessness. Third, /a/ identification rates (blue dashed line) fall the most dramatically and are the lowest in tokens where the medial target vowel is spliced out, suggesting that only high vowels are potentially available for recovery.

Because the results of splice-1 and splice-3 tokens show no surprising trends, the rest of this paper will focus on the splice-0 (full-vowel), splice-2 (no vowel), and splice-4 (no vowel and no C_1 burst/frication) results. The splice-2 results will also be compared against naturally produced vowel-less tokens to test how the presence of coarticulatory cues affect the responses.

3.1. Tokens with full medial vowel

Shown below in Figure 4 are vowel identification rates for tokens with a full target vowel of 40 ms, broken down by context and by C_1. The figure shows that the identification rates for /i, a/ are essentially at ceiling, but identification rates for /u/ are surprisingly low at below 90%.

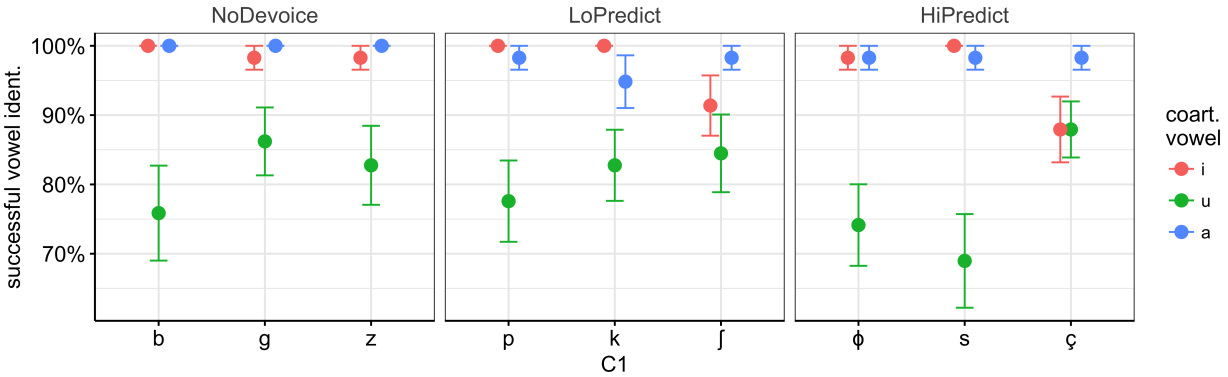

Figure 4: Successful vowel identification in VC_1VC_2V tokens with full medial vowel.

The most common wrong response by the participants for [u] identification was ∅ for all C_1 as shown in Figure 5, meaning that the participants either heard the vowel accurately or confused [u] with ∅, but rarely confused the vowel with another vowel. The confusion specifically between $/C_1C_2/$ and $/C_1uC_2/$ sequences suggests two things. First, $[C_1uC_2]$ can be mapped to both $/C_1uC_2/$ and $/C_1C_2/$, although there is a bias towards the former. Second, the fact that there is confusion between $/C_1uC_2/$ and $/C_1C_2/$ even when a vowel of 40 ms is fully present suggests that the distinction between the two sequences is weak, and that C_1C_2 and C_1uC_2 sequences are treated as more or less equivalent by Japanese listeners. While this provides some support for the account presented by Dupoux and colleagues, the participants also exhibit some confusion between /i/ and ∅ after /ʃ, ç/, which most likely stems from /i/ being the most common vowel after these consonants, which was shown in Table 3.

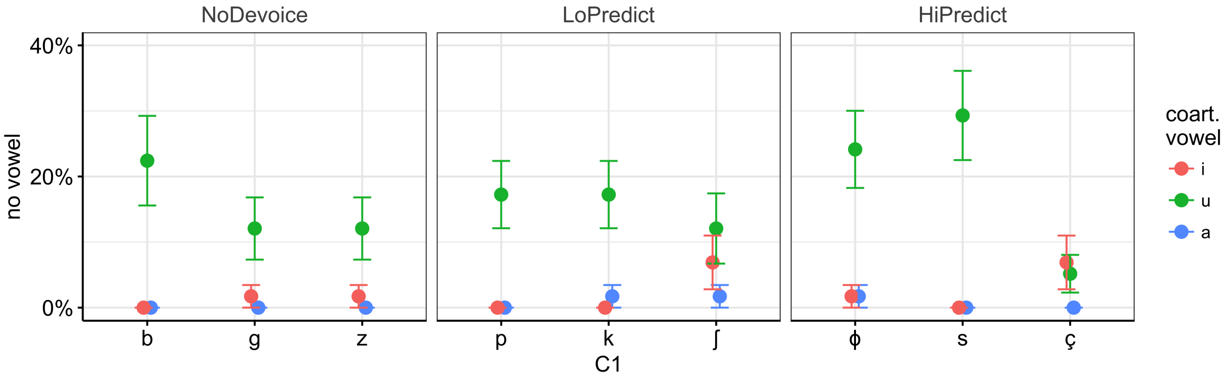

Figure 5: "No vowel" responses for VC_1VC_2V tokens with full medial vowel.

The results of the full-vowel tokens suggest that stimuli such as [epko, epu̥ko, epuko] are possibly all being treated as equivalent to /epuko/. Because they all map to the same phonotactically legal structure, there is bidirectional repair, although with a bias towards vowel recovery. The fact that there is confusion for /u/ across the board, even for /g/ despite /a/ being the most common vowel to follow, provides some support to the phonetically minimal repair hypothesis presented by Dupoux and colleagues. However, the fact that there is also confusion for /i/ after /ʃ, ç/ additionally suggests that phonotactic probability affects perception as well.

3.2. Tokens with no medial vowel

This section compares the results of naturally vowel-less tokens and the splice-2 tokens where the medial, phonated vocalic material has been completely removed but C_1 burst/frication noise fully remains. Acoustically, the difference between these tokens is that the naturally vowel-less tokens contain no obvious coarticulatory information, unlike the spliced tokens.

3.2.1. Naturally vowel-less tokens

The prediction in terms of vowel detection was that the rate of <∅> responses should be highest for non-devoicing contexts since high vowel devoicing is rare in these contexts making Japanese listeners more sensitive to the presence versus absence of a medial vowel. Conversely, the rate of <∅> responses was expected to be low in contexts where high vowel devoicing is expected, since high vowels can delete in these contexts, leading to a bias towards recovery. This bias should be especially high in HiPredict contexts because one of two high vowels can be predicted with near-absolute certainty, leading Japanese listeners to disregard phonetic cues that may contradict their contextual vowel prediction.

Presented first below in Table 4 are the responses for naturally produced VCCV tokens. Bold numbers indicate the most frequent responses for a given C_1. A chi-square test was performed using the *chisq.test()* function in R to test whether the observed response rates were significantly different from chance. /a/ responses were excluded under the assumption that /a/ is not a candidate for recovery and also because /a/ responses were at or near 0% in most contexts. The results showed that the observed responses were significantly different from chance at $p < 0.01$ with the exception of /p/ ($p = 0.4909$).

Table 4: Responses for naturally produced VC_1C_2V tokens. Most frequent responses in bold.

	NoDevoice			LoPredict			HiPredict		
	ebko	egto	ezpo	epko	ekto	eʃpo	eɸko	espo	eçto
a	0.14	0.02	0.03	0.10	0.02	0.00	0.00	0.00	0.00
i	0.10	0.05	0.09	0.24	0.02	**0.55**	0.07	0.07	**0.76**
u	0.34	0.43	**0.50**	0.29	**0.59**	0.26	**0.60**	**0.60**	0.14
∅	**0.41**	**0.50**	0.38	**0.36**	0.38	0.19	0.33	0.33	0.10

Overall, the results show that <∅> responses are 50% or lower across all contexts, revealing an overall bias towards perceptual repair. However, the rate of <∅> responses is highest for NoDevoice environments as predicted, suggesting that there indeed is an effect of high vowel devoicing. Additionally, <∅> responses are lowest for HiPredict environments, confirming the prediction that predictability has an effect on the rate of repair as well.

The responses to naturally vowel-less tokens also suggest that there is an effect of phonotactics that drives the choice of vowel that is recovered by Japanese listeners. The vowel recovered after [ʃ, ç] is, again, /i/ rather than /u/, further strengthening the account that the choice of the vowel used for phonotactic repair is not just merely a default, minimal vowel but rather chosen based on phonotactic probability. This is also in line with a recent finding by Durvasula and Kahng (2015), who also found in Korean listeners that the choice of recovered vowel is better predicted by the phonological alternations observed in the language rather than a phonetically minimal repair strategy.

3.2.2. Spliced vowel-less tokens (Splice-2)

Shown below in Figure 6 are vowel identification rates for tokens with all of the vowel spliced out, broken down by context and C_1. The figure shows that the identification rates for high vowels are highest for low-predictability contexts but remain above 40% in the other two contexts. Identification rates for /a/ is generally lower than for high vowels across all contexts, but is clearly lowest in high-predictability contexts.

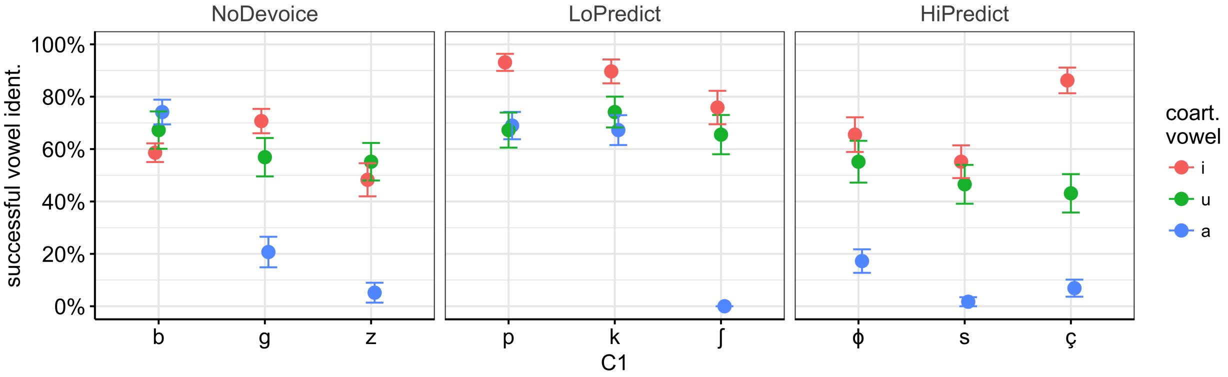

Figure 6: Successful identification rate of target vowel for spliced VCVCV tokens.

Another prediction going into the current study was that participants should be most sensitive to high vowel coarticulatory cues in contexts where high vowel devoicing is expected, especially in low-predictability contexts. A mixed logit model was fit using the *glmer* function of the *lme4* package of R, with successful vowel identification rates as the dependent variable. The statistical analysis compares the rate of correct identification of spliced vowels from coarticulatory cues, so naturally produced VCCV tokens, which should contain no vowel coarticulatory cues, are not included in the analysis. The fixed effects structure of the model consisted of target vowel, context, and their interaction. Because target vowel and context are categorical variables with more than two levels, a deviation coding scheme was used so that each level is compared to the grand mean of the rate of vowel identification for all vowels across all contexts (Clopper, 2013). The model with a fully-crossed, maximal random effects structure failed to converge, hence the final random effects structure included by-participant and by-stimulus random intercepts as well as by-participant random slopes for target vowel. The interaction was shown to be a non-significant contributor to the fit of the model ($p = 0.4546$), and thus was excluded from the final model. The results are shown below in Table 5. The diacritic for devoicing (i.e., V̥) is used throughout to indicate the vowels that have been spliced out.

Table 5: Mixed logit model results comparing successful vowel identification rates across difference predictability contexts. Compared against grand mean.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	0.13765	0.23562	0.584	0.55909	
[ḁ]	-1.67112	0.32931	-5.075	3.88e-07	***
[i̥]	1.31109	0.32505	4.033	5.50e-05	***
[u̥]	0.3600	0.3508	1.026	0.30480	
NoDevoice	-0.05901	0.29508	-0.200	0.84151	
LoPredict	0.85552	0.29667	2.884	0.00393	**
HiPredict	-0.7965	0.2969	-2.683	0.00730	**

The results confirm the prediction. The rate of successful vowel identification was significantly higher than or similar to the grand mean for high vowels ([i̥, u̥], respectively) but significantly lower for [ḁ]. Context also showed significant effects, where the rate of successful vowel identification was significantly higher in LoPredict contexts. Vowel identification was also significantly worse in HiPredict contexts.

3.2.3. Comparison of naturally vowel-less and spliced vowel-less tokens

Naturally vowel-less tokens and spliced tokens by themselves tell only part of the story. Another prediction was that Japanese listeners should be able to recover high vowels from the coarticulatory information in spliced tokens, leading to differences between splice-2 and naturally vowel-less tokens. If it is the case that phonotactic violation alone is responsible for vowel epenthesis and that the choice of vowel is the phonetically minimal segment, namely /u/, then the presence of vowel coarticulatory information should do little to affect the choice of vowel.

Shown in Table 6 below are the results of a mixed logit model that compares detection rates for spliced tokens compared to the grand mean of naturally vowel-less tokens across all contexts. The results show that [i̥] coarticulation but not [u̥] and [ḁ] coarticulation drives up the vowel responses significantly. Additionally, the significant effects of context suggest that Japanese listeners are less likely to perceptually epenthesize a vowel in NoDevoice contexts and more likely in HiPredict contexts.

Table 6: Mixed logit model results comparing vowel detection between VCCV and spliced VCVCV tokens.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	0.8535	0.2007	4.253	2.11e-05	***
[i̥]	1.2575	0.2028	6.201	5.61e-10	***
[u̥]	0.2540	0.1974	1.286	0.198323	
[ḁ]	0.2264	0.1664	1.361	0.173505	
NoDevoice	-0.5232	0.1444	-3.623	0.000292	***
LoPredict	0.1057	0.1587	0.666	0.505455	
HiPredict	0.4176	0.1556	2.683	0.00729	**
[i̥]:NoDevoice	0.2076	0.2349	0.884	0.376749	
[u̥]:NoDevoice	0.3405	0.2105	1.618	0.105718	
[ḁ]:NoDevoice	0.4924	0.2036	2.418	0.015585	*
[i̥]:LoPredict	0.4977	0.2656	1.874	0.06093	.
[u̥]:LoPredict	0.1221	0.2159	0.566	0.57162	
[ḁ]:LoPredict	0.4701	0.2155	2.182	0.02913	*
[i̥]:HiPredict	-0.7053	0.2403	-2.935	0.00333	**
[u̥]:HiPredict	-0.4626	0.2157	-2.144	0.03201	*
[ḁ]:HiPredict	-0.9625	0.2058	-4.677	2.91e-06	***

For identification rates, spliced tokens are compared separately to naturally vowel-less tokens to make the effects of coarticulation for each vowel clearer. Presented below in Table 7 below are the responses for spliced [u] tokens and Figure 7 shows how the responses compare to naturally vowel-less tokens.

Table 7: Responses for $VC_1(u)C_2V$ tokens with medial vowel spliced out. Most frequent response in bold.

	NoDevoice			LoPredict			HiPredict		
	ebu̥ko	egu̥to	ezu̥po	epu̥ko	eku̥to	eʃu̥po	eɸu̥ko	esu̥po	eçu̥to
a	0.02	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.02
i	0.02	0.00	0.12	0.00	0.00	0.09	0.03	0.05	**0.47**
u	**0.67**	**0.57**	**0.55**	**0.67**	**0.74**	**0.66**	**0.55**	0.47	0.43
∅	0.29	0.43	0.31	0.33	0.26	0.26	0.40	**0.48**	0.09

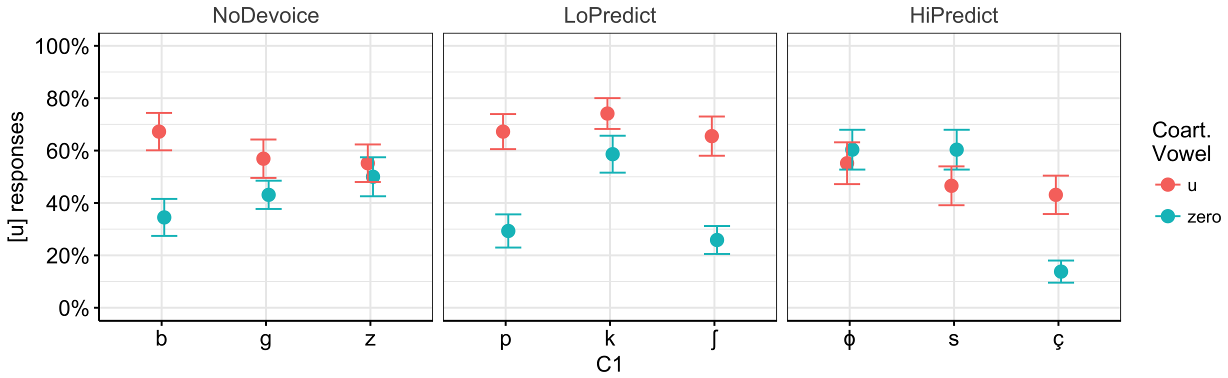

Figure 7: <u> responses for naturally vowel-less vs. spliced [u] tokens.

A mixed logit model was fit to the data with the rate of <u> responses as the dependent variable. <u> was chosen since it is regarded as the default epenthetic segment. The predictors were stimulus type (i.e., spliced vs. natural), C_1, and their interactions. C_1 was used as a predictor rather than context because the epenthetic vowel does not seem to be uniform across all contexts but rather depend on C_1. By-participant and by-stimulus random intercepts were included. By-participant random slopes for target vowel and C_1 were also included. All predictors were significant contributors to the fit of the model. The model results are shown below in Table 8, with the grand mean of ∅ tokens (i.e., naturally vowel-less) tokens across all C_1 as the baseline.

Table 8: Mixed logit model results comparing <u> responses between VCCV and spliced VC(u)CV tokens.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	-0.4373	0.2341	-1.868	0.06174	.
[u̥]	1.0252	0.2188	4.685	2.8e-06	***
[b]	-0.3954	0.3585	-1.103	0.26997	
[g]	0.1085	0.3383	0.321	0.74843	
[z]	0.4425	0.3516	1.259	0.20814	
[p]	-0.7999	0.3759	-2.128	0.03336	*
[k]	0.9138	0.3711	2.462	0.01380	*
[ʃ]	-0.8408	0.4023	-2.090	0.03661	*
[ɸ]	1.1924	0.4427	2.693	0.00707	**
[ç]	-1.7374	0.5285	-3.287	0.00101	**
[s]	1.1127	0.4032	2.760	0.00579	**
[b]:[u̥]	0.8645	0.5049	1.712	0.08685	.
[g]:[u̥]	-0.3378	0.4765	-0.709	0.47830	
[z]:[u̥]	-0.7469	0.4688	-1.593	0.11108	
[p]:[u̥]	1.3755	0.5304	2.594	0.00950	**
[k]:[u̥]	0.2426	0.5329	0.455	0.64895	
[ʃ]:[u̥]	1.1375	0.5063	2.247	0.02466	*
[ɸ]:[u̥]	-1.3770	0.5311	-2.593	0.00952	**
[ç]:[u̥]	0.8267	0.5327	1.552	0.12069	
[s]:[u̥]	-1.9827	0.5241	-3.783	0.00016	***

The model shows that the presence of a coarticulated vowel does indeed significantly raise the overall rate of <u> responses compared to naturally vowel-less tokens. The model also shows clearly that none of the consonants from the NoDevoice context nor their interaction with [u] coarticulation have a significant effect on the overall rate of <u> responses. A separate model was fit just for the NoDevoice consonants, and the results showed that [u] coarticulation raised the overall rate of <u> responses (p = 0.00565), but this raising effect was driven by [b] tokens, where [bu̥] tokens had significantly higher <u> responses than [b∅] tokens (p = 0.0036). The other two consonants showed no significant effect of [u] coarticulation.

In the case of LoPredict consonants, [p, ʃ] drive down the rate of <u> responses for naturally vowel-less tokens, and the interaction terms show that this consonantal effect is significantly mitigated in [u] spliced tokens. This means that the high rate of <u> responses for spliced tokens remain high even when the preceding consonants are [p, ʃ]. A separate model was fit for the LoPredict consonants, and the results showed that [u] coarticulation raised the overall rate of <u> responses (p = 0.0011). [pu̥, ʃu̥] tokens had significantly higher <u> responses than [p∅, ʃ∅] tokens (p = 0.0435, 0.0445, respectively), but [k] tokens showed no significant difference.

The overall model also shows that the HiPredict consonants [ɸ, s] drive up the rate of <u> responses in naturally vowel-less tokens and also that this raising effect is mitigated for [u] spliced tokens. A separate model was also fit for the HiPredict consonants, and the results showed that vowel coarticulation did not have a significant effect overall (p = 0.55256). Only [ç] tokens showed a significant effect, where [çu̥] tokens had significantly higher <u> responses than [ç∅] tokens (p = 0.0051).

<i> responses were also driven up by [i] coarticulation compared to the baseline of naturally vowel-less tokens. Shown below in Table 9 is a summary of the responses for spliced [i] tokens and Figure 8 shows how the rate of <i> responses compare to naturally vowel-less tokens.

Table 9: Responses for $VC_1(i)C_2V$ tokens with medial vowel spliced out.

	NoDevoice			LoPredict			HiPredict		
	ebi̥ko	egi̥to	ezi̥po	epi̥ko	eki̥to	eʃi̥po	eɸi̥ko	esi̥po	eçi̥to
a	0.09	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	**0.59**	**0.71**	**0.48**	**0.93**	**0.90**	**0.76**	**0.66**	**0.55**	**0.86**
u	0.10	0.10	0.41	0.03	0.03	0.10	0.19	0.24	0.03
∅	0.22	0.16	0.10	0.03	0.07	0.14	0.16	0.21	0.10

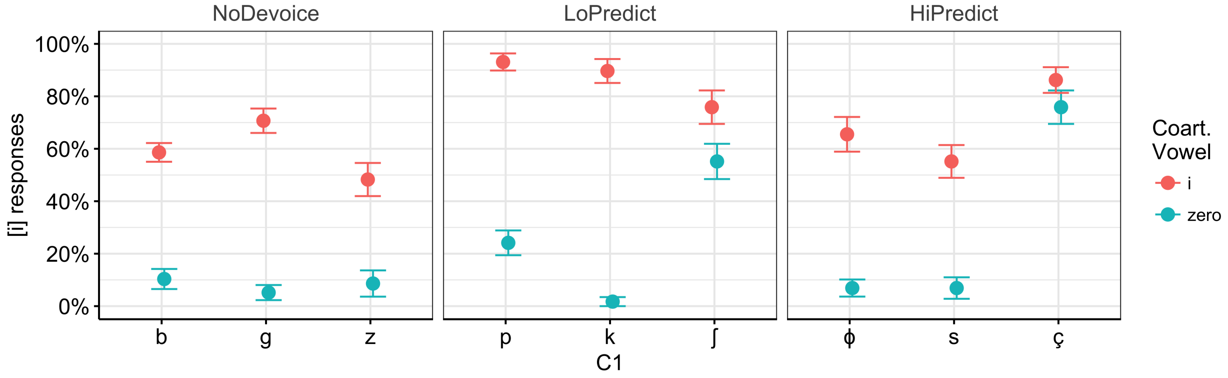

Figure 8: <i> responses for naturally vowel-less vs. spliced [i] tokens.

As was the case with [u] tokens, vowel coarticulation seems to affect which vowel participants report to hearing. A similar model as in Table 8 was fit, with the same predictors and random effects structure. The dependent variable was <i> responses with naturally vowel-less tokens as the baseline. The results are shown in Table 10 below.

Table 10: Mixed logit model results comparing <i> responses between VCCV and spliced VC(i)CV tokens.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	-3.049661	0.516151	-5.908	3.45e-09	***
[i̥]	5.2063	0.632616	8.230	< 2e-16	***
[b]	0.0033	1.015582	0.003	0.997405	
[g]	-0.8108	1.125833	-0.720	0.471406	
[z]	-0.3929	1.050555	-0.374	0.708387	
[p]	0.8335	1.045082	0.798	0.425107	
[k]	-5.8767	2.929376	-2.006	0.044841	*
[ʃ]	3.4272	0.993605	3.449	0.000562	***
[ɸ]	-1.0112	1.140228	-0.887	0.375143	
[ç]	4.5973	0.985998	4.663	3.12e-06	***
[s]	-0.7629	1.08516	-0.703	0.482018	
[b]:[i̥]	-1.2371	1.381969	-0.895	0.370692	
[g]:[i̥]	0.1918	1.453363	0.132	0.894990	
[z]:[i̥]	-1.8775	1.369134	-1.371	0.170284	
[p]:[i̥]	0.8714	1.447218	0.602	0.547110	
[k]:[i̥]	10.879399	3.309833	3.287	0.001013	**
[ʃ]:[i̥]	-3.8467	1.318740	-2.917	0.003535	**
[ɸ]:[i̥]	0.05834	1.433339	0.041	0.967532	
[ç]:[i̥]	-3.9666	1.340264	-2.960	0.003081	**
[s]:[i̥]	-1.0877	1.40003	-0.777	0.437199	

Again, [i] coarticulation has a raising effect on the rate of <i> responses, but none of the NoDevoice consonants show significant effects. A separate model was fit to these consonants, and the results showed that [i] coarticulation raised the overall rate of <i> responses (p < 0.0001). All spliced tokens had significant higher <i> responses than naturally vowel-less tokens as well (p < 0.0001).

The overall results also showed that [ʃ, ç] drive up the rate of <i> responses in naturally vowel-less tokens, which is unsurprising, since /i/ is the most probably vowel after these consonants. Separate models for LoPredict and HiPredict contexts were fit, and the LoPredict results showed that the difference between spliced and naturally vowel-less tokens were significant overall (p < 0.001). The raising effect of [i], however, was not significant for [ʃ] (p = 0.9998). The HiPredict results mirrored that of the LoPredict model in that the raising effect of [ç] in naturally vowel-less tokens was large enough to make the difference between spliced and naturally vowel-less tokens non-significant (p = 0.5608).

Thus far, the results suggest that the choice of epenthetic vowel for Japanese listeners is not simply a default /u/, but rather that the choice of vowel is sensitive to the acoustic cues in the signal. /u, i/ are both high vowels that are targeted for devoicing in Japanese, so this is perhaps not surprising. Japanese listeners have had a lifetime of practice attending to subtle coarticulatory cues to recover devoiced and deleted high vowels. Then what about a vowel like /a/, which rarely undergoes devoicing? The responses to spliced [a] tokens are shown in Table 11 below and Figure 9 shows how the responses compare between naturally vowel-less and spliced tokens.

Table 11: Responses for $VC_1(a)C_2V$ tokens with medial vowel spliced out.

	NoDevoice			LoPredict			HiPredict		
	ebḁko	egḁto	ezḁpo	epḁko	ekḁto	eʃḁpo	eɸḁko	esḁpo	eçḁto
a	**0.74**	0.21	0.05	**0.69**	**0.67**	0.00	0.17	0.02	0.07
i	0.00	0.03	0.09	0.00	0.00	**0.57**	0.10	0.03	**0.52**
u	0.12	*0.38*	**0.55**	0.09	0.19	0.26	0.21	0.47	0.28
∅	0.14	*0.38*	0.31	0.22	0.14	0.17	**0.52**	**0.48**	0.14

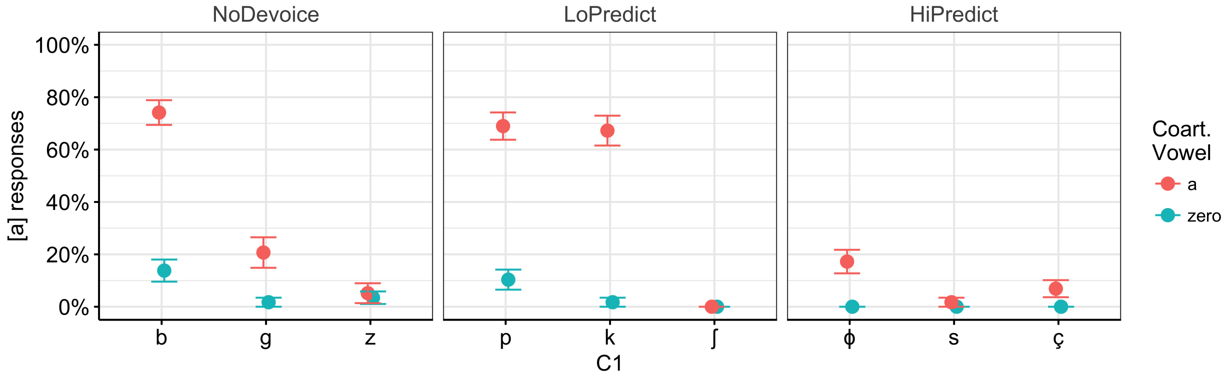

Figure 9: <a> responses for naturally vowel-less vs. spliced [a] tokens.

Although limited to post-stop environments (i.e., [b, g, p, k]), the results show that participants can recover the spliced [a] vowel at relatively high rates. Bilabial place also seems to have a facilitatory effect. Given that <a> responses were generally low in naturally vowel-less tokens, the raising effect even in the limited environments is surprising. The fact that /a/ identification is limited to stops may be due to the articulatory differences between stops and fricatives. Because stops have a portion in which there is no airflow, coarticulation with the following vowel can be more complete by the time the stop burst/aspiration occurs. This is also true of bilabial place, where the lack of lingual gesture allows the following vowel to be coarticulated earlier. This is less true of fricatives where the transition into a fricative is more gradual, and coarticulation with the following vowel occurs towards the end of the segment rather than throughout. Since /a/ is a low vowel that a Japanese listener does not often have to recover, it may be that the beginning of the fricative already leads to the listener anticipating a high vowel and ignore the low vowel cue towards the end.

A mixed logit model with the same predictors and random effects structure as in Tables 8 and 10 was fit. Responses to [ʃ] tokens were removed from the model because /a/ responses were at 0% for both the naturally vowel-less and spliced [a] tokens, resulting in no meaningful difference. When included in the model, [ʃ] tokens had an extremely low intercept of -17, but an absurdly high standard error of 6,999, both of which are most likely errors stemming from an absolute lack of difference between participants. The interaction between target vowel and C$_1$ was not a significant contributor to the fit of the model and thus was excluded in the final model. The results are shown below in Table 12. The dependent variable was <a> responses with the grand mean of naturally vowel-less tokens across all C$_1$ as the baseline.

Table 12: Mixed logit model results for <a> responses. [ʃ] excluded.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	-7.878226	0.877458	-8.978	< 2e-16	***
[ḁ]	4.596604	0.675904	6.801	1.04e-11	***
[b]	4.796960	0.810236	5.920	3.21e-09	***
[ɡ]	0.006499	0.858052	0.008	0.99396	
[z]	-4.369956	1.657550	-2.636	0.00838	**
[p]	4.665762	0.871624	5.353	8.65e-08	***
[k]	3.646247	0.805332	4.528	5.96e-06	***
[ɸ]	1.032369	1.009703	1.022	0.30657	
[ç]	-0.847829	1.033016	-0.821	0.41180	
[s]	-6.9843	2.0675	-3.378	0.00073	***

The results confirm that indeed [a] coarticulation does have a significant raising effect on the rate of <a> responses. The bilabial stops [b, p] have a significant raising effect on the rate of <a> responses, while the alveolar fricatives [z, s] have a significant lowering effect.

The models above all compared spliced tokens to naturally vowel-less tokens to investigate whether perceptual repairs by Japanese listeners are automatic as previously claimed and thus treated as the same. First, the vowel detection results reveal that Japanese listener are more likely to perceive a vowel in devoicing contexts than in non-devoicing contexts. In high predictability contexts, where vowel detection is highest, the perceived vowel is the most probable high vowel (i.e., /ɸu, su/ and /çi/). Because the detection rates are already high for naturally vowel-less tokens, the presence of coarticulatory cues did not significantly change the rate of vowel responses. In low predictability contexts, coarticulation effects on the rate of vowel identification were significant for both high vowels with the exception of /ʃi/ tokens. There was also an unexpected effect of [a] coarticulation, but this was limited to stop contexts. Together, the detection and identification results suggest that there is indeed a bias towards repairing phonotactically illegal consonant clusters, but the epenthetic vowel is chosen due to a combination of phonotactic predictability and sensitivity to phonetic cues.

3.3. Tokens with no vowel and no burst/short frication noise

The results discussed in §3.2 for spliced vowel-less but burst-ful tokens (splice-2) show that Japanese listeners are biased towards perceiving a vowel between heterorganic consonant clusters, and that the choice of vowel is sensitive to the coarticulatory cues present in the C$_1$ burst/frication noise. Numerous studies have shown that the presence of a stop burst or frication noise in phonotactically illegal sequences are often interpreted as signaling the presence of a vowel (see Davidson & Shaw, 2012, Hsieh, 2013 for English; Furukawa, 2009, Whang, 2016 for Japanese; Kang, 2003 for Korean). This section therefore discusses the results of splice-4 tokens, where the target vowel has been spliced out completely and C_1 also has been spliced out leaving just the closure for stop C_1 and <15 ms of frication noise for fricative C_1.

The responses to all splice-4 tokens are summarized in Table 13 below. A mixed logit model was fit to test whether the rates of <∅, i, u, a> responses were significantly affected by the identify of the vowel that was spliced out. Stop C_1 and fricative C_1 were analyzed separately. The results revealed that the responses were not significantly different regardless of the target vowel, with the exception of spliced [u̥] tokens where C_1 was /b/, which drove up <u> responses (p = 0.002333). Because the effect was limited to a single consonant, this section collapses the responses across all target vowels and focuses more on vowel detection.

Table 13: Responses for VC_1˺C_2V tokens with medial vowel and C_1 burst/frication noise spliced out.

	NoDevoice			LoPredict			HiPredict		
	eb˺ko	eg˺to	ez˺po	ep˺ko	ek˺to	eʃ˺po	eɸ˺ko	es˺po	eç˺to
a	0.05	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.01
i	0.08	0.13	0.16	0.07	0.02	0.36	0.09	0.10	0.45
u	0.32	0.17	0.34	0.07	0.09	0.11	0.11	0.14	0.10
∅	0.55	0.68	0.47	0.85	0.87	0.52	0.78	0.76	0.45

The results show first and foremost that the rate of <∅> responses never reaches 100%. This is perhaps expected for fricative C_1, since there was ∼15 ms of frication remaining in the tokens. Factors contributing to the results for stop C_1, on the other hand, are less obvious. A mixed logit model was fit separately for the stops and fricatives since the the fricative tokens had a short frication noise remaining whereas the stop tokens had no burst at all. The full model for both data subsets had the following structures. The fixed effects included context, V_1, and their interaction. All stimuli used in the experiment had the form $\mathrm{V}_1\mathrm{C}_1(\mathrm{V})\mathrm{C}_2\mathrm{V}_2$, where the order of V_1-V_2 was always either [e-o] or [o-e]. V_1 was included as a predictor to test whether the ordering of the initial and final vowels had a significant effect on vowel detection, which would suggest that there might be V-to-V coarticulatory cues that the participants are picking up on. The random effects included by-participant and by-stimulus random intercepts as well as by-participant random slopes for context, V_1, and their interaction.

Shown first below in Table 14 is the result of the final model for the stop-only subset. Since the HiPredict context had no stops, the subset only includes NoDevoice and LoPredict contexts with the latter as the baseline. The interaction was shown to be a non-significant contributor to the fit of the model (p = 0.5463) and thus was removed.

Table 14: Mixed logit model result for vowel detection in spliced vowel-less and burst-less stop tokens.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	-1.8118	0.3162	-5.730	1.00e-08	***
V_1 = [o]	-0.4413	0.3369	-1.310	0.19	
NoDevoice	1.5019	0.3493	4.299	1.71e-05	***

The results show that V_1 did not have a significant effect, but the rate of vowel detection was significantly higher for NoDevoice tokens than LoPredict tokens. A possible explanation for this effect is that the C_1 in NoDevoice tokens had consistent phonation during closure, as shown in Figure 10 below.

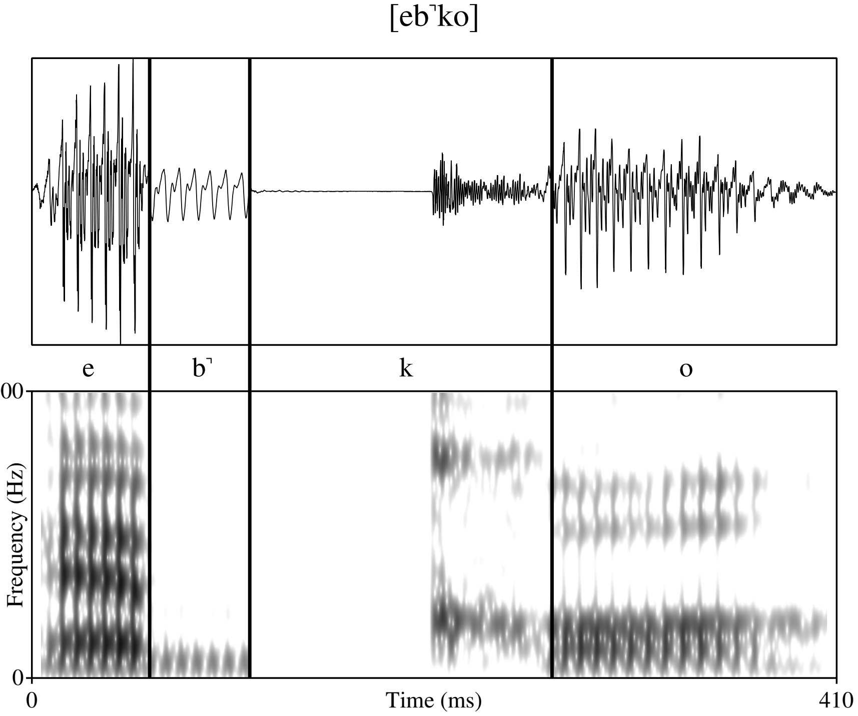

Figure 10: Spliced vowel-less, burst-less token created from [ebako].

The mixed logit model for the fricative-only subset also showed that the vowel detection rate for the NoDevoice fricative [z] is significantly higher than for HiPredict fricatives although not higher than the LoPredict fricative [ʃ] (p = 0.658). For the fricatives, only context was a significant contributor to the fit of the model, and thus V_1 (p = 0.81919) and V_1:Context (p = 0.82666) were excluded from the fixed effects structure of the final model. The results are shown below in Table 15.

Table 15: Mixed logit model result for vowel detection in splice-4 fricative tokens.

	Estimate	Std. Error	z	Pr(>|z|)	
(Intercept)	-0.9063	0.3023	-2.998	0.00272	**
LoPredict	0.7666	0.5488	1.397	0.16243	
NoDevoice	1.0569	0.5065	2.087	0.03691	*

Although the fact that vowel detection rates never fall to 0% can be easily explained by the presence of prevoicing for NoDevoice tokens and the 15 ms frication noise for the fricatives, the 10+% of vowel detection for the LoPredict stops [p˺, k˺] is still somewhat puzzling. Without a vowel and without a burst between C_1 and C_2, a token such as [ep˺ko] contains a doubly long stop closure, much like a geminate medial consonant as in [ekko]. Geminate consonants are phonotactically legal in Japanese and require no repair. Nevertheless, participants report perceiving a vowel some of the time. It is possible that some participants are picking up on the mismatch between the transitional cues out of V_1 and into V_2. This seems unlikely, however, in that transitional cues into a vowel often outweighs transitional cues out of a vowel for Japanese listeners (Fujimura et al., 1978) and that Japanese listeners rely more on centroid spectral cues than on formant transitions (Hirai et al., 2005). Perhaps a more likely explanation is one of task effect. Although the stimuli sounded as though they contain a geminate obstruent, there was no geminate option given as a possible answer. This might have kept the participants from fully eliminating the vowel-ful answer choices, and having been exposed to numerous vowel-ful tokens (both acoustically and perceptually) during the task, the participants might have assumed that a vowel should be present at least some of the time.

3.4. Summary of main findings

There were five main findings in the perceptual experiment. First, Japanese listeners seem to sometimes confuse the high vowel that is phonotactically the most likely after a given C_1 with ∅ even when the high vowel is 40 ms long and fully phonated. This sort of confusion was not observed with the low vowel /a/, which is typically not devoiced. Second, results from naturally vowel-less tokens revealed that the vowel most often perceptually epenthesized between illicit clusters is /u/, largely due to the fact that it is phonotactically the most probable vowel after most obstruents in Japanese. This is further supported by the finding that after /ʃ, ç/, which is most often followed by /i/ rather than /u/, the choice of epenthetic vowel is in fact /i/. Third, participants successfully identified spliced high vowels in splice-2 tokens (full C_1 with target vowel completely spliced out) at rates significantly higher than the baseline rates observed in naturally vowel-less tokens. Identification rates of spliced /a/ were significantly lower and limited to after stops. Fourth, related to the third finding, identification rates of high vowels were lowest in HiPredict contexts, suggesting that listeners are less sensitive to low-level coarticulatory cues in contexts where the phonotactics typically is sufficient for identifying the target vowel. Lastly, <∅> responses never quite reach 100% even for splice-4 tokens where both C_1 and target vowel were fully spliced out suggesting a bias towards CV structure, but this may have also been due to task effects.

4. Discussion

The aim of the current study was to test whether Japanese listeners utilize coarticulatory cues more in contexts where the listener is expected to have less certainty regarding their predicted vowel (low predictability) than in contexts where the vowel can be predicted with high certainty (high predictability). Broadly speaking, overt consonant clusters were shown to be mapped to a phonotactically legal CVC sequence, neutralizing the contrast between CC and CVC sequences as Dupoux and colleagues have shown shown. However, the specific vowel recovered was shown to be modulated by CV co-occurrence probabilities in Japanese and that CC and CVC were not perceived as equivalent. This difference in CC and CVC processing in Japanese listeners was also found in Cutler et al. (2009).

First, the perception of full-vowel tokens showed that there is confusion between /u/ and ∅, even when there is a 40 ms-long, phonated [u]. It is possible that this confusion arises because /u/ is indeed the default epenthetic vowel in Japanese, making it equivalent to ∅. However, a survey of biphone co-occurrence probabilities in the Corpus of Spontaneous Japanese revealed that /u/ also happens to be the most common vowel after most consonants, making it difficult to attribute the default status of /u/ as stemming simply from its shortness (Dupoux et al., 1999, 2011). Furthermore, similar confusion with ∅ is observed for /i/ after /ʃ, ç/, suggesting that the choice of epenthetic vowel must be conditioned by the phonotactic probabilities of the language.

Second, the perception of vowel-less tokens further suggests that Japanese listeners confuse vowel-ful and vowel-less tokens with a tendency towards vowel-fulness. The results for splice-4 (vowel-less and burst-less) tokens in particular showed that Japanese listeners interpret even the most minute acoustic cues such as prevoicing of stops as signaling the presence of a vowel (§3.3). However, participants do not seem to simply perceive a default vowel. A comparison between naturally vowel-less and spliced vowel-less tokens showed that spliced tokens drive up the rate of target vowel responses significantly. This suggests that while heterorganic C_1C_2 sequences are perceived as being equivalent to C_1VC_2 as Dupoux and colleagues argue, the particular vowel is again not simply the “default” but the result of sensitivity to the acoustic information in the signal as dictated by the listener’s native language. The participants, therefore, are recovering the vowel that is the most probable based on the phonetic cues contained in the burst/frication noise of C_1.

Third, the rate of high vowel identification was above chance at 40% across all contexts in spliced vowel-less tokens. Specifically, recovery rates were the highest in LoPredict contexts as predicted, and the recovery rates were significantly lower for HiPredict contexts, also as predicted. Recovery rates in NoDevoice contexts fell somewhere between the two devoicing contexts. The high rates of recovery suggest that Japanese listeners are hypersensitive to vowel coarticulatory cues, and the lower rate of recovery in HiPredict contexts additionally suggests that sensitivity to coarticulatory cues are conditioned by phonotactic predictability.

Lastly, sensitivity to coarticulatory cues in Japanese listeners is limited primarily to high vowels.

The participants were worst at identifying /a/. Non-high vowels are typically not devoiced in Japanese, and thus Japanese listeners have relatively little experience recovering them.

5. Conclusion

Based on the results discussed above, perhaps the terms perceptual epenthesis and “illusory” vowel epenthesis should be not used interchangeably. The “default” vowel is not [u] in Japanese simply because it is the shortest, but because it is the most common high vowel that Japanese listeners have been trained to recover all their lives. Phonotactic repair in Japanese listeners, therefore, is more akin to perceptual repair, where they use phonotactic and phonetic processes to choose the most probable vowel. In contrast, Brazilian Portuguese lacks a similar systematic devoicing process, and thus phonotactic repairs by Brazilian Portuguese listeners as reported by Dupoux et al. (2011) might be more “illusory” in nature, triggered primarily by phonotactic violations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1524133.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

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

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

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Clopper, C. G. (2013). Modeling multi-level factors using linear mixed effects. In *Proceedings of meetings on acoustics* (Vol. 19, p. 060028).

Cutler, A., Otake, T., & McQueen, J. M. (2009). Vowel devoicing and the perception of spoken Japanese words. *Journal of the Acoustical Society of America*, *125*(3), 1693–1703.

Davidson, L., & Shaw, J. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics*, *40*(2), 234-248.

Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, *12*, 635-647.

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

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*, *64*, 199-210.

Durvasula, K., & Kahng, J. (2015). Illusory vowels in perceptual epenthesis: The role of phonological alternations. *Phonology*, *32*(3), 385-416.

Ernestus, M. (2011). Gradience and categoricality in phonological theory. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (p. 2115-36). Wiley-Blackwell.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults' perception of / r/ and / l/. *Journal of the Acoustical Society of America*, *99*, 1161-1173.

Fujimoto, M. (2015). Vowel devoicing. In H. Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology* (chap. 4). Mouton de Gruyter.

Fujimura, O., Macchi, M., & Streeter, L. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, *21*(4), 337-346.

Furukawa, K. (2009). *Perceptual similarity in loanword adaptation between Japanese and Korean* (Unpublished master's thesis). University of Toronto.

Hall, K. C., Hume, E., Jaeger, F., & Wedel, A. (in preparation). *The message shapes phonology.* (Ms. University of British Columbia, University of Canterbury, University of Rochester & Arizona University.)

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *Acoustical Society of America*, *96*, 73-82.

Hirai, S., Yasu, K., Arai, T., & Iitaka, K. (2005). Acoustic cues in fricative perception for Japanese native speakers. *Technical Report of Institute of Electronics, Information and Communication Engineers*, *104*(696), 25-30.

Hsieh, C.-H. (2013). *The perception of epenthetic vowels in voiced and voicless contexts in Japanese* (Unpublished master's thesis). University of Kansas.

Hume, E., Johnson, K., Seo, M., Tserdanelis, G., & Winters, S. (1999). A cross-linguistic study of stop place perception. In *Proceedings of the 14th International Congress of Phonetic Sciences* (p. 2069-2072).

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*(29), 9541-9546.

Kang, Y. (2003). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology*, *20*(2).

Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language*, *83*(2), 536–574.

Kong, E. J., Beckman, M., & Edwards, J. (2012, November). Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *Journal of Phonetics*, *40*(6), 725-744.

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

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*((1-2)), 71-102.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477-500.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language and Linguistics*, *16*(6), 791-818.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., … Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*, 432-434.

Norris, D. (1994). Shortlist: A connectionist model of con- tinuous speech recognition. *Cognition*, *52*, 189-234.

Pierrehumbert, J. (2001). Exemplar dynamics: word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (p. 137-157). Amsterdam: John Benjamins.

Pitt, M., & McQueen, J. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*, 347–370.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Shademan, S. (2006). Is phonotactic knowledge grammatical knowledge? In D. Baumer, D. Montero, & M. Scanlon (Eds.), *Proceedings of the 25th west coast conference on formal linguistics* (p. 371-379).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379-423.

Shaw, J., & Kawahara, S. (2018). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics*, *66*, 100-119.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273-293.

Vance, T. (1987). *An Introduction to Japanese Phonology.* New York: SUNY Press.

Varden, J. K. (2010, March). Acoustic correlates of devoiced Japanese vowels: velar context. *The Journal of English and American Literature and Linguistics*, *125*, 35-49.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science*, *9*, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374-408.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*, 47–62.

Whang, J. (2016). Perception of illegal contrasts: Japanese adaptations of Korean coda obstruents. In *Proceedings of Berkeley Linguistics Society* (Vol. 36).

Whang, J. (2018). Recoverability-driven coarticulation: Acoustic evidence from Japanese high vowel devoicing. *Journal of the Acoustical Society of America*, *143*(2), 1159-1172.

Wilson, C., Davidson, L., & Martin, S. (2014). Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language*, *77*, 1-24.