Article

Language *and* Speech

Auditory Contrast versus Compensation for Coarticulation: Data from Japanese and English Listeners

Language *and* Speech
54(4) 499–525

DOI: 10.1177/0023830911404959
las.sagepub.com
SAGE

John Kingston, Shigeto Kawahara, Daniel Mash and Della Chambless
University of Massachusetts, USA

Abstract
English listeners categorize more of a [k-t] continuum as "t" after [ʃ] than [s] (Mann & Repp, 1981). This bias could be due to compensation for coarticulation (Mann & Repp, 1981) or auditory contrast between the fricatives and the stops (Lotto & Kluender, 1998). In Japanese, surface [ʃk, ʃt , sk, st] clusters arise via palatalization and vowel devoicing from /sik, sit, suk, sut/, and acoustic vestiges of the devoiced vowels remain in the fricative. On the one hand, compensation for coarticulation with the devoiced vowel would cancel out compensation for coarticulation with the fricative, and listeners would not show any response bias. On the other hand, if the stop contrasts spectrally with the fricative, listeners should respond "t" more often after [ʃi̥] than [su̥]. Experiment 1 establishes that [k] and [t] coarticulate with preceding voiced [i, u], voiceless [i̥, u̥], and [ʃ, s]. Experiment 2 shows that both Japanese and English listeners respond "t" more often after [ʃi̥] than [su̥], as predicted by auditory contrast. English listeners' "t" responses also varied after voiced vowels, but those of Japanese listeners did not. Experiment 3 shows that this difference reflects differences in their phonetic experience.

Keywords
auditory contrast, compensation for coarticulation, crosslinguistic comparison, English, Japanese

Introduction

Neighboring sounds affect the identification of speech sounds. Mann and Repp (1981) found that English listeners identify a stop that is ambiguous between [k] and [t] more often as "t" after [ʃ] than after [s]. Mann and Repp propose that listeners hear a stop articulated after [ʃ] as being pronounced further back than after [s], perceptually undo this backing, and hear the stop as "t". In this

Corresponding author:
John Kingston, Linguistics Department, University of Massachusetts, 150 Hicks Way, 226 South College, Amherst, MA 01003-9274, USA
Email: jkingston@linguist.umass.edu

explanation, listeners compensate for coarticulation. Alternatively, listeners may hear the spectrum of a stop ambiguous between [k] and [t] as higher in frequency—more like [t]—after the relatively low spectral concentration of energy in the fricative [ʃ]. In this explanation, the target contrasts auditorily with its context (see Lotto & Kluender, 1998, for the first proposal along these lines).

Whether these response biases are byproducts of compensation for coarticulation or auditory contrast remains an open question (see Diehl, Lotto, & Holt, 2004; Fowler, 2006; Lotto & Holt, 2006, for recent overviews), largely because both accounts predict that the target will be heard as different from its context. They differ only in whether the perceived difference is articulatory or auditory.

The current study takes advantage of a difference between the phonologies of Japanese and English that may permit a choice between these alternatives. The Japanese sibilant fricative is pronounced as [s] before the vowel [u], but as [ʃ] before [i]. Furthermore, the two high vowels devoice and disappear as distinct acoustic intervals between two voiceless consonants. Devoicing thus creates the surface clusters [ʃk, ʃt, sk, st] from /sik, sit, suk, sut/ sequences (Han, 1962; McCawley, 1977; Beckman, 1982; Beckman & Shoji, 1984; Tsuchida, 1994, 1997).

Vestiges of devoiced vowels nevertheless remain. Beckman and Shoji (1984) argue that speakers still produce the vowel's oral articulation, even though they fail to adduct the glottis. Nakamura (2003) finds traces of the devoiced vowels' oral articulations in his electropalatographic data. The higher formants produced by the vowel articulation can be seen in a spectrogram during the interval of fricative noise (Han, 1962; Figure 4 below), and the fricative noise is longer when it includes a devoiced vowel (Han, 1994). Japanese listeners even perceive a vowel between heterorganic consonants when none is there because the phonotactics of the language prohibits such clusters (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Dupoux, Pallier, Kakehi, & Mehler, 2001). Although tangible traces of devoiced vowels remain and Japanese phonotactics encourages listeners to perceive the vowel as still present, the differences in the place of articulation of the fricative—[s] when the vowel is [u̥] and [ʃ] when it is [i̥]—are large and categorical, and may therefore be more noticeable than the vestiges of the devoiced vowels.

These facts lead to two competing hypotheses concerning how listeners would categorize a [k-t] continuum following [su̥] and [ʃi̥]:

1. Hypothesis 1: Listeners hear the stops' spectra as contrasting auditorily with the preceding fricatives' spectra.
2. Hypothesis 2: Listeners compensate for coarticulation with the voiceless vowel and the fricative.

If the stop's spectrum contrasts auditorily with its context's spectrum, then the predominant spectral properties of the context—the energy concentration in the fricative—should bias listeners to respond "t" more often after [ʃi̥] than [su̥]. Alternatively, listeners could compensate for coarticulation with both the fricative and the voiceless vowel, because both articulations occur in the noise interval, and the stop coarticulates with both. Because the voiceless vowel [i̥] would pull the stop forward while [ʃ] would pull it back, these two biases should cancel one another out, and there would be no effect of context.

If the compensatory adjustments cancel one another out, then the compensation hypothesis predicts no contextual effects, while the contrast hypothesis predicts a positive effect. Both hypotheses predict, however, that voiced vowels will bias listeners' responses.

Experiment 1 consists of an acoustic analysis of coarticulation between [k] and [t] and preceding voiced and voiceless vowels and fricatives in Japanese. Experiment 2 shows how these contexts affect Japanese and English listeners' categorization of a [k-t] continuum. Experiment 3 tests the effects of a larger array of preceding voiced vowels on Japanese and English listeners' response biases.

2 Experiment 1: Coarticulation of stops with preceding voiced and voiceless vowels and fricatives in Japanese

To clarify the prediction of the compensation for coarticulation account, Experiment 1 measured coarticulation between the stops and preceding fricatives in [ʃk, ʃt , sk, st] clusters derived from /sik, sit, suk, sut, sjuk, sjut/ as well as between [k, t] and preceding voiced vowels [i, u] in Japanese.

2.1 Method

2.1.1 Speakers. Two male and two female adult speakers of Tokyo Japanese were recorded—the second author was one of the male speakers. None reported any hearing or speaking disorder. The recording sessions lasted 20 minutes, and the three outside recruits were paid $5. All participants in this and the other two experiments gave informed consent before participating.

2.1.2 Materials. The speakers each produced 10 randomized repetitions of nonce words consisting of three syllables: (1) [ɾu], (2) {na, nu, ni, sʉ̥, ʃi̥, ʃʉ̥},[1] and (3) {to, ko} in the frame *dʒaa* ___ *de onegai* "Please do X with ___" (these stimuli mimic the properties of the stimuli used for Experiment 2). The stimuli were pronounced with the initial low-high F0 contour of unaccented words (Haraguchi, 1977). The speakers were encouraged to speak casually, which ensured that all the high vowels were devoiced after the voiceless fricatives. The output of the head-mounted microphone worn by the speakers was digitized at 44100 kHz with 16 bit resolution. They repeated the tokens 10 times in a randomized order. Only the last 9 repetitions were analyzed.

2.1.3 Measurements. To assess the coarticulatory interactions between [k, t] and preceding segments, three values were measured: (1–2) the spectral center of gravity and standard deviation of the stop burst of [k] and [t], and (3) F2 at the onset of voicing in the following vowel. Compared to [t], [k] is expected to have a lower center of gravity and a smaller standard deviation and to be followed by a lower F2 at the onset of the vowel [o] (Jakobson, Fant, & Halle, 1952; Stevens & Blumstein, 1978; Blumstein & Stevens, 1979). Coarticulation with preceding fricatives differing in place and vowels differing in backness is expected to alter each of these correlates of stop place.

All measurements and calculations were made in Praat (Boersma & Weenink, 2007). The spectral center of gravity and standard deviation were calculated from an FFT of a 25 ms long interval centered on the burst, within a Gaussian window, band-pass filtered between 750–6000 Hz (roughly the range of F2–F6). The stops' mean VOT exceeded 20 ms, so this interval included little if any of the voiced portion of the vowel. F2 at the onset of voicing was extracted using Praat's LPC algorithm, with the number of formants set to 5, and the maximum frequency of the formants set to 5000 Hz for the male speakers and to 5500 Hz for the female speakers.

2.1.4 Statistical analyses. Three types of analyses were conducted. The independent variables in each analysis were: (1) voiced [ni, nu] versus voiceless [ʃi̥, sʉ̥] and front [ni, ʃi̥] versus back [nu, sʉ̥], (2) voiced [ni, nu] versus voiceless [ʃi̥, ʃʉ̥] and front [ni, ʃi̥] versus back [nu, ʃʉ̥], and (3) front [sʉ̥] versus back [ʃʉ̥]. The second analysis unconfounded vowel backness from fricative place, while the third unconfounded fricative place from vowel backness. All analyses included two other independent variables, speaker and stop place (k/t).[2]

2.2 Results

2.2.1 Center of gravity of the stop burst's spectrum. Figure 1a first shows that the differences between [k]'s and [t]'s centers of gravity differed more after [u] than [i] for both voiced and voiceless vowels

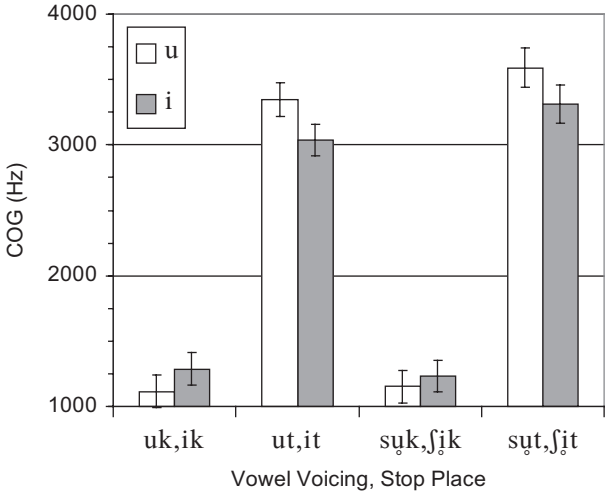

Figure 1a. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [u, i]. Voiceless [u̥] and [i̥] are simultaneous with [s] and [ʃ], respectively.

(white versus gray bars). Second, [t]'s but not [k]'s center of gravity was lower after voiced than voiceless vowels (second versus fourth pair of bars). Voice was significant, $F(1, 269) = 7.289$, $p = .007$, but vowel was only marginally significant, $F(1, 269) = 3.217$, $p = .074$, and both variables interacted significantly with stop place, voice by stop place: $F(1, 269) = 8.175$, $p = .005$; vowel by stop place: $F(1, 269) = 19.902$, $p < .001$. The three-way interaction was not significant, $F < 1$.

Figure 1b shows that when fricative place is held constant, voiceless [i̥] lowers both [k]'s and [t]'s centers of gravity compared to voiceless [u̥] (gray versus white bars). Voiced [i] shrinks the difference in center of gravity between [k] and [t] compared to voiced [u] (gray versus white bars in the first and second pairs of bars). Voice was still significant, $F(1, 277) = 4.041$, $p = .045$, but vowel was no longer even marginally significant, $F(1, 277) = 1.505$, $p > .10$. Vowel nonetheless interacted significantly with stop place, $F(1, 277) = 4.125$, $p = .043$, and the three-way interaction between voice, vowel, and stop place was significant, $F(1, 277) = 9.942$, $p < .001$.

Finally, Figure 1c shows that the difference in center of gravity between [k] and [t] was larger after [sụ̥] than [ʃụ̥] (white versus gray bars). Fricative place did not significantly affect the center of gravity of the stop burst's spectrum on its own ($F < 1$), but did interact significantly with stop place, $F(1, 134) = 4.973$, $p = .027$.

Figure 1a shows that the difference in center of gravity between [k] and [t] was larger after [sụ̥] than [ʃi̥]. Figure 1c shows the same difference, and moreover that the apparent effect of the voiceless vowels' backness in Figure 1a was the influence of the fricative's place. Comparing the voiced vowels' effects in Figure 1b with the fricatives' effects in Figure 1c shows that voiced [u] and [s] both increased the size of difference in the center of gravity between [k] and [t], compared to voiced [i] and [ʃ].

Contrary to expectation, neither [u] and [s] nor [i] and [ʃ] have opposite effects on the spectral center of gravity of the following stop's burst. The more posterior sounds in each pair, [u] and [ʃ], were expected to pull both the stops' articulations back, and by lengthening the cavity in front of the constriction, they should lower the centers of gravity in the burst spectra compared to the more anterior sounds. However, there is no contradiction in the front voiced vowel [i] and the back

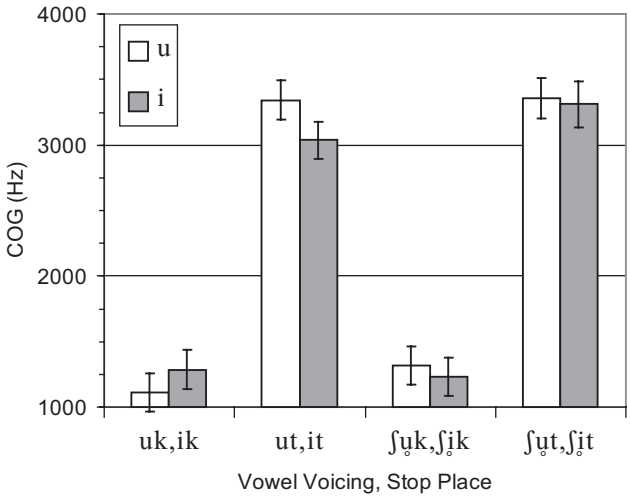

Figure 1b. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [ʃ].

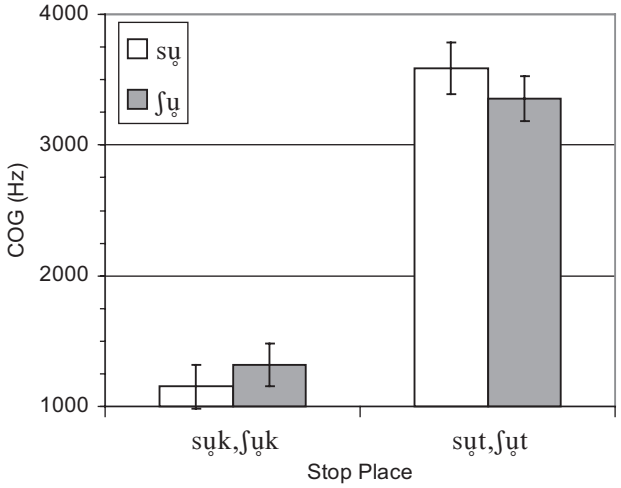

Figure 1c. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiceless [u̥]. Voiceless [u̥] is simultaneous with both [s] and [ʃ].

fricative [ʃ] having the same coarticulatory effects on the stop burst's center of gravity. The lingual constrictions in [i] and [ʃ] are adjacent—palatal and palato-alveolar—and both fall between [k]'s and [t]'s constriction locations, so coarticulation with either one should pull [k] forward and [t] backward, and thereby shrink the difference between their centers of gravity. While coarticulation with [u] would pull [t] backward and coarticulation with [s] would pull [k] forward, coarticulation with [u] keeps [k] back and coarticulation with [s] keeps [t] front. As a result, center of gravity differences between [k] and [t] do not shrink in these contexts compared to after [i] and [ʃ]. This parallelism between fricative and vowel contexts is striking because [u] is the vowel that causes the

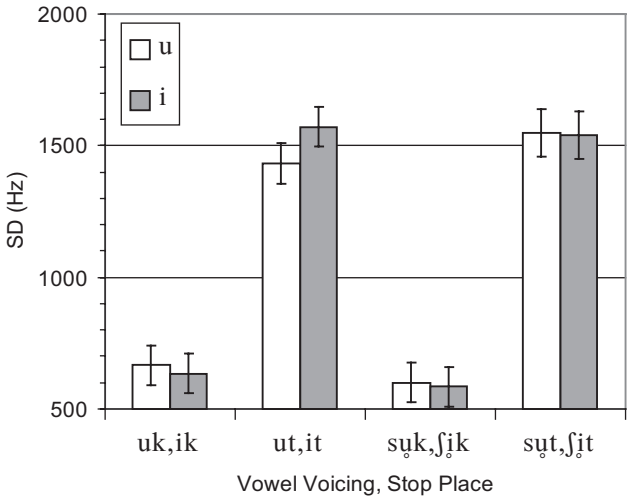

Figure 2a. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [u, i]. Voiceless [u̥] and [i̥] are simultaneous with [s] and [ʃ], respectively.

Japanese sibilant fricative to be pronounced [s], and [i] is the vowel that causes it instead to be pronounced [ʃ].

This analysis thus predicts that a listener would compensate for coarticulation in the same way after [u] as [s] and in the same way after [i] as [ʃ], and that compensating for the difference between the vowels should not cancel out the effect of compensating for the difference between the fricatives (cf. Hypothesis 2). However, this prediction applies only to the voiced vowel contexts, and not the voiceless ones, too, as no reliable difference in [k]'s and [t]'s centers of gravity was obtained when the preceding vowels were voiceless. That outcome predicts that listeners would instead only compensate for coarticulation with the preceding fricatives.

2.2.2 Standard deviation of the stop burst's spectrum. Figure 2a shows that the standard deviation was larger for diffuse [t] than compact [k] (second and fourth pairs of bars versus the first and third pairs), but otherwise differed little between the contexts [ni, nu, ʃi̥, sụ̥]. Neither voice nor vowel was significant alone, both Fs < 1, nor were their interactions with stop place any more than marginally significant, voice by stop place: $F(1, 269) = 3.127$, $p = .078$; vowel by stop place: $F(1, 269) = 2.367$, $p > .10$; voice by vowel by stop place: $F(1, 269) = 2.127$, $p > .10$.

Figure 2b shows that the difference between [k] and [t] was larger after [i] than [u] (gray versus white bars), and that the voiceless vowels affected this difference at least as much as the voiced vowels (third and fourth pairs of bars versus the first and second pairs). Voice was just marginally significant, $F(1, 277) = 2.779$, $p = .097$, and did not interact with stop place, $F < 1$, but vowel did reach significance, $F(1, 277) = 4.458$, $p = .036$, and interacted significantly with stop place, $F(1, 277) = 15.557$, $p < .001$. The three-way interaction was not significant, $F(1, 277) = 1.208$, $p > .10$.

Figure 2c shows that the difference between [k] and [t] was larger after [sụ̥] than [ʃụ̥] (white versus gray bars). Fricative place was only marginally significant, $F(1, 134) = 3.097$, $p = .081$, but interacted significantly with stop place, $F(1, 134) = 9.592$, $p = .002$. [s] and [ʃ] have the opposite effect on the standard deviation of the stop burst's spectrum from the vowels that determine their place of articulation (cf. Figures 1a, 1b).

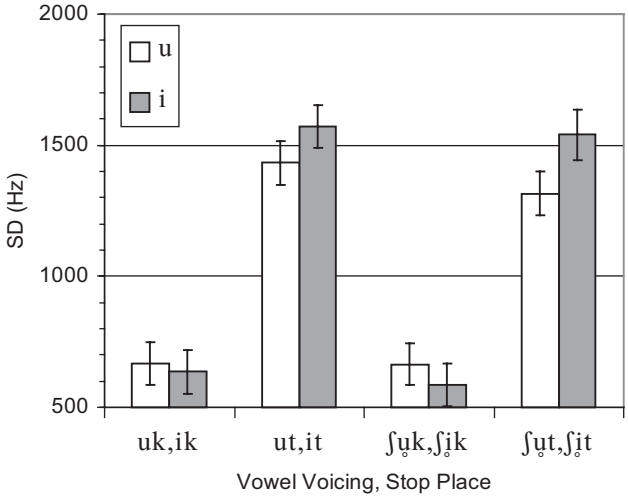

Figure 2b. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [ʃ].

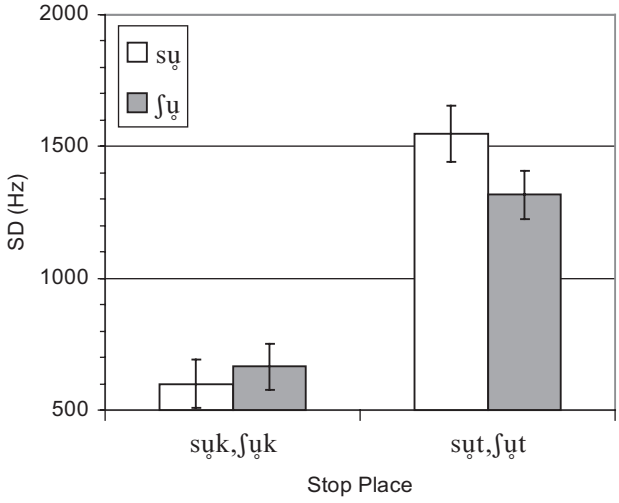

Figure 2c. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiceless [u̥]. Voiceless [u̥] is simultaneous with both [s] and [ʃ].

By pulling the [t] constriction back, [ʃ] makes its spectrum less diffuse, and by pulling the [k] constriction forward it makes its spectrum less compact. Because [s] is articulated as far forward as [t], coarticulation with [s] does not alter the diffuseness of [t]'s spectrum, while [s] does pull [k] forward and make its spectrum less compact. The result is that the difference between the standard deviations of [t]'s and [k]'s spectra does not contract as much after [s] as [ʃ].

On the one hand, coarticulation with the front vowel [i] would pull [k]'s constriction toward the front of the palate; as a result, F3 and F4 would be drawn closer together. Coarticulation with [i]

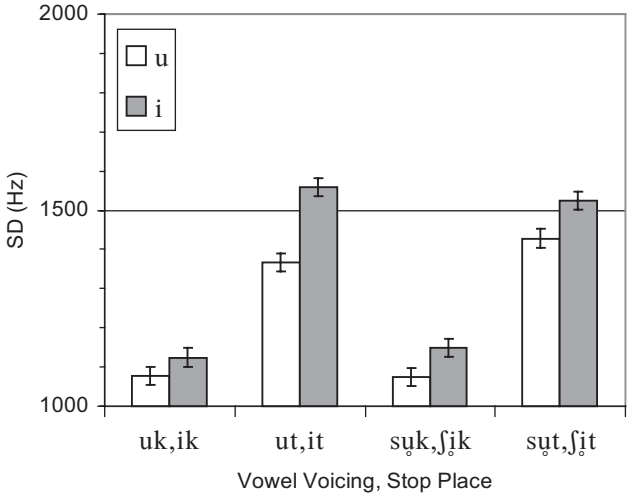

Figure 3a. Mean F2 values (95% confidence intervals) at voice onset following [k] vs. [t], preceded by voiced and voiceless [u, i]. Voiceless [ụ̥] and [i̥] are simultaneous with [s] and [ʃ], respectively.

would not alter [t]'s articulation much, so these formants would remain well separated. On the other hand, coarticulation with the back vowel [u] would pull [k]'s articulation back, which would draw F2 and F3 closer together. Coarticulation with [u] would also pull [t]'s articulation backward and thereby make its spectrum less diffuse. The result is that the compact–diffuse difference between [k]'s and [t]'s burst spectra is greater after [i] than [u].

By the standard deviation measure, stops do coarticulate with preceding voiceless vowels as well as voiced ones. Moreover, the acoustic effects of coarticulation with both voiced and voiceless vowels are opposite those of coarticulation with the fricatives, a result which predicts that compensating for coarticulation with one context would cancel out the compensation for coarticulation with the other. That fricative place has the opposite effect from vowel backness on the standard deviations of stops' burst spectra also explains why no consistent effect of context was obtained in the analysis that confounds these two variables.

2.2.3 Second formant frequency at the onset of voicing in the following [o]. Figure 3a shows that voiceless vowels affect F2 following [k] more than voiced vowels (third versus first pair of bars). However, voiced vowels have the bigger effect on following [t] (second versus fourth pair). Nonetheless, values were consistently higher after [i] than [u] for both [k] and [t]. Vowel was significant, $F(1, 281) = 152.42$, $p < .001$, voice was not, $F(1, 281) = 2.109$, $p > .10$, and they interacted significantly with one another, $F(1, 281) = 4.198$, $p = .041$, and with stop place, $F(1, 281) = 13.614$, $p < .001$.

Figure 3b shows that voiceless vowels exerted a smaller coarticulatory effect than voiced ones, especially for [k] (third and fourth pairs of bars versus the first and second pairs). Vowel and voice were both significant, $F(1, 286) = 76.403$, $p < .001$, $F(1, 286) = 17.248$, $p < .001$, they interacted significantly, $F(1, 286) = 22.192$, $p < .001$, and both interacted significantly with stop place, $F(1, 286) = 5.378$, $p = .021$.

Figure 3c shows that F2 is higher after [ʃụ̥] that [sụ̥]. Fricative place was significant, $F(1, 145) = 14.413$, $p < .001$, but it did not interact significantly with stop place, $F(1, 145) = 2.35$, $p > .10$.

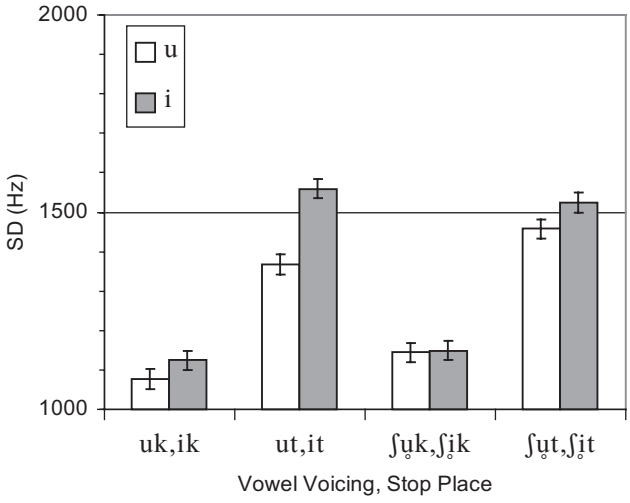

Figure 3b. Mean F2 values (95% confidence intervals) at voice onset following [k] vs. [t], preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [ʃ].

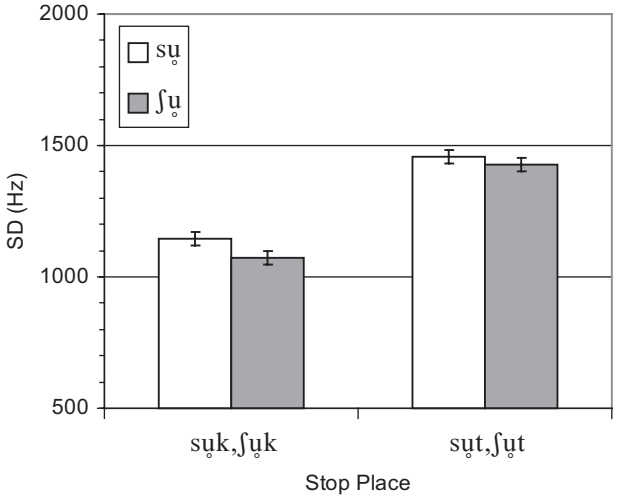

Figure 3c. Mean F2 values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiceless [u̥]. Voiceless [u̥] is simultaneous with both [s] and [ʃ].

Fricative place has the same effect as the vowel backness differences that determine fricative place: F2 is higher after [ʃ] than [s] and higher after [i] than [u].

Higher F2 onset frequencies are expected following [i] than [u], and likewise following [ʃ] than [s] (Mann & Repp, 1980; Whalen, 1981; Nittrouer, Studdert-Kennedy, & McGowan, 1989; Jongman, Wayland, & Wong, 2000). If listeners compensate for these acoustic effects, their stop place judgments would be biased in the same direction after both [i̥] and [ʃ] and likewise after both [u̥] and [s]. However, compensation for these coarticulatory effects would lead them to respond "t" **less** often after [i̥] and [ʃ] than after [u̥] and [s].

2.3 Discussion

Experiment 1 found that a preceding fricative's place of articulation significantly affected all three measures of coarticulation. Stop burst spectra's centers of gravity differed as a function of a preceding **voiced** vowel's backness, but not a preceding voiceless vowel's. The standard deviations of the stop's burst spectra, however, differed at least as much after voiceless as voiced vowels, and F2's onset frequency likewise differed after voiceless as well as voiced vowels. These results show that listeners have good acoustic evidence to compensate for coarticulation with preceding fricatives, and likewise preceding vowels, whether they are voiceless or voiced.

The bursts' centers of gravity and the following F2's onset frequency were altered in the same direction by both [i, i̥] and [ʃ] and in the opposite direction by both [u, u̥] and [s]. Listeners might therefore be expected to compensate in the same way for coarticulation with the vowels as with the fricative in each pair—contrary to the prediction in Hypothesis 2. However, compensation for the effects of coarticulation on the spectra's centers of gravity would induce more "t" responses after [i, i̥, ʃ] than [u, u̥, s], while compensation for its effects on F2 onset frequency in the following vowel would induce fewer "t" responses in the former contexts than the latter.

Because [i, i̥] and [ʃ] have similar constriction locations and active articulators, they should alter the articulation of following stops similarly, and listeners could readily use the acoustic consequences of these alterations to compensate for their alteration of its articulation. Neither the constriction location nor the active articulator is similar in [u, u̥] and [s]. Because the mapping from acoustics to articulations is one-to-many for [u, u̥] and [s], the acoustic consequences of coarticulation do not provide listeners with information that is specific to a unique articulatory source. Attributing those consequences to [u, u̥]'s back articulation would pull the stop percept forward, attributing them instead to [s]'s front articulation would pull its percept back, and attributing them to both segments would leave the stop percept unchanged because the two compensatory adjustments would cancel each other out.

Compensation for the coarticulatory effects on the standard deviation of the stop bursts' spectra leads to yet different predictions. The standard deviation differs from the other two acoustic measures in being altered in opposite directions by [i, i̥] versus [ʃ] and [u, u̥] versus [s], and compensating for one coarticulatory effect would cancel out compensating for the other, as predicted by Hypothesis 2. Cancellation is predicted to be entirely general here, in that compensating for coarticulation with both [i, i̥] and [ʃ] cancels one another's effects just as much as compensating for coarticulation with both [u, u̥] and [s].

We next report the results of an experiment that tests the predictions of the competing hypotheses regarding the perception of stop place in these contexts.

3 Experiment 2: Response biases induced by preceding voiced vowels and fricatives

3.1 Method

3.1.1 Stimuli. The stimuli were three-syllable nonce words of the form [ɾuʃi̥Co] or [ɾusu̥Co], where C represents a stop drawn from a [k-t] continuum. Mid-back rounded [o] serves as the ideal vowel to follow the target stop because other vowels either would devoice in that position and affricate the [t], or would cause the stimuli to resemble existing words. Furthermore, the transitional probabilities between [k] and [t] and a following [o] are more similar than for any other following vowel (Amano & Kondo, 2000).

Table 1. Values (Hz) of F1–F4 at the end of the vowels used as stimuli in Experiments 2 and 3.

Vowel	F1	F2	F3	F4
i	317	2449	3128	3824
u	315	1367	2707	3363
a	710	1622	2840	4159
e	537	1885	2679	3859
o	534	1062	2353	3374

The initial syllable [ɾu] staves off lexical effects, as *suto* "strike" and *niko* "two pieces" are real Japanese words. The materials were also synthesized with the unaccented initial low-high F0 contour (Haraguchi, 1977) to ensure that they would be perceived as nonce words—real bimoraic loanwords, such as *suto* and *niko*, have an accented initial high-low contour. These efforts were successful in that no Japanese listeners recognized any of our stimuli as real words.

To test the spectral effects of fully voiced vowel contexts, stimuli of the form [ɾuniCo], [ɾunaCo], and [ɾunuCo] were also used. Both the auditory contrast and compensation for coarticulation accounts predict that English and Japanese listeners should respond "t" least often after voiced [i], more often after voiced [a], and most often after voiced [u].

3.1.2 Recording. A native speaker of Tokyo Japanese (the second author) pronounced nonce words of the form [ɾunV$_i$hV$_i$], where V$_i$ = {i, a, u}, and [ɾuSV$_j$hV$_j$] where SV$_j$ = {ʃi̥, su̥}. The final syllable consisted of [h] followed by a copy of the second vowel so that the second vowel was not altered by coarticulation with any following consonant nor the vowel in the next syllable. These utterances were produced in the frame *dʒaa* ___ *de onegai* "Please do X with ___" in a sound-attenuated recording booth and digitized at a sampling rate of 44100 Hz with 16 bit resolution.

3.1.3 Synthesis and stimulus construction. F0, formant, and intensity contours extracted from the sonorant intervals of representative tokens were used as parameters for resynthesizing these intervals via the Sensyn implementation of KLSYN88 (Klatt & Klatt, 1990). Table 1 lists values of F1–F4 at the end of each of the three vowels (the values listed for [e, o] are for stimuli used in Experiment 3).

Naturally produced [ʃi̥] and [su̥] were used for fricatives: the [ʃi̥] had the lowest frequency energy concentration of 10 recorded tokens, while the [su̥] had the highest (Figure 4). The third formant of devoiced [i̥] can be seen in the spectral peak at about 3000 Hz in Figure 4a, and the second formant of devoiced [u̥] is visible in the spectral peak at about 1800 Hz in Figure 4b.

Naturally produced [k] and [t] bursts taken from recordings of [ɾuhoko] and [ɾuhoto] were used as endpoints for a 12-step stop place continuum. Their durations were trimmed to 15 ms, their intensities were adjusted, and they were mixed with inversely varying intensities to form a continuum roughly half of which was heard as "k" and the other half as "t". The onset frequencies of the [o]'s formants were also manipulated to convey this place difference. Values defining the formant trajectories of the endpoint stimuli are listed in Table 2. Intermediate steps were interpolated linearly between these endpoint values. The silent interval simulating the stop closure lasted 40 ms following voiced vowels, and 60 ms following voiceless syllables.[3]

The acoustic results of these manipulations are displayed in Figures 5–8. Figure 5 shows spectrograms of four complete stimuli, [ɾunako, ɾunato, ɾuʃi̥ko, ɾusu̥ko]. Figure 6 shows spectrograms of the beginnings of the [ko] and [to] endpoints of the [k-t] continuum and the stop bursts' spectra. F2 starts higher at the [to] than the [ko] endpoint. Energy is distributed across a broad range of

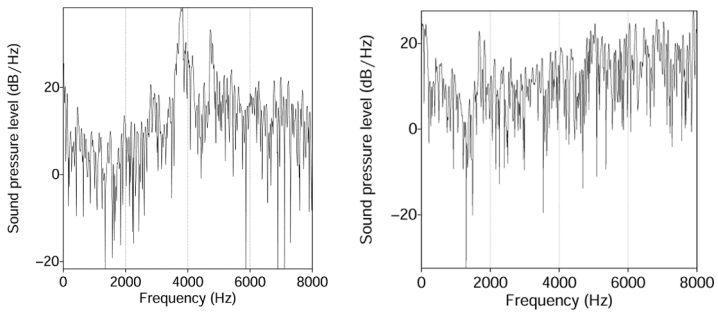

Figure 4. Spectra of (a) [ʃi̥] and (b) [su̥] showing their energy distributions between 0–8000 Hz. The spectra were calculated from a 50 ms wide Gaussian window centered in each fricative.

Table 2. Time (ms) and value (Hz) pairs for F1–F5 at the [k] and [t] endpoints used in synthesizing the [o] in the third syllable of the stimuli for Experiments 2 and 3.

F1		Time	F2		F3		F4		F5	
ms	Hz	ms	k	t	k	t	k	t	k	t
0	275	0	900	1450	2400	2800	3300	3800	4100	4600
70	475	25	900	1450	2400	2800	3300	3800	4100	4600
150	475	45	860	973	2600	2600	3480	3480	4193	4193
		150	850	850	2600	2600	3480	3480	4193	4193

frequencies in [t]'s stop burst, while in [k]'s burst, energy is instead concentrated in a single peak just below 1000 Hz. Figure 7 shows spectrograms of the voiced vowel [ni, na, nu] contexts and spectra of the latter part of the vowels (top to bottom panels). There is a broad peak in [i]'s spectrum above 2000 Hz, comprising F2–F4 (top panel), [a]'s spectrum has a prominent peak between 1000–2000 Hz, consisting of F2 (middle panel), and the bulk of the energy in [u]'s spectrum is below 1000 Hz, consisting of F1 alone because F2 is quite weak (bottom panel). Finally, Figure 8 shows spectrograms of the [ʃi̥] and [su̥] intervals, and spectra from late in these intervals. Energy levels rise with increasing frequency in [su̥]'s spectrum, particularly above 4000 Hz, but in [ʃi̥]'s spectrum energy is concentrated in a comparatively narrow peak between 3000–4000 Hz.

These variations create the necessary conditions for listeners to perceive the targets as contrasting spectrally with the preceding contexts. The contexts also differ categorically in both vowel and consonant place, there are acoustic vestiges of the devoiced vowels in the fricative intervals, and the targets vary between two endpoints that also differ in place, so these contexts likewise create the necessary conditions for listeners to compensate for coarticulation of the targets with these contexts.

3.1.4 Participants. Twenty-one native speakers of Japanese were recruited in Tokyo. They received 1000 yen per hour for their time. Data from one listener who gave "ko" responses to nearly all the stimuli were excluded from analysis. None of them participated in Experiment 1. Twenty native speakers of English were recruited from the University of Massachusetts, Amherst community. They earned either course credit or $10/hour for their participation. All participants were older than

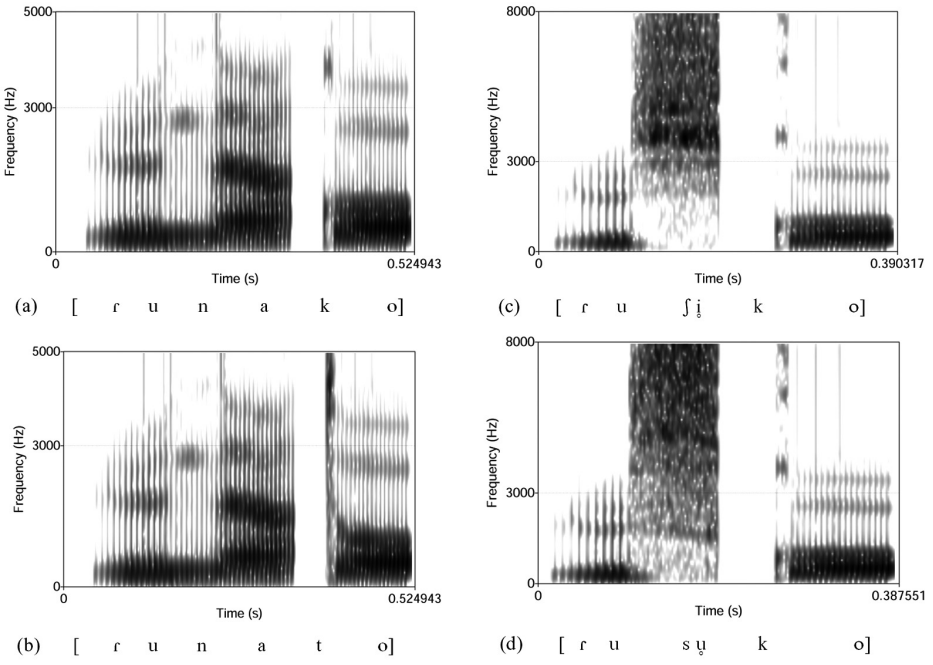

Figure 5. Spectrograms of example stimuli: (a) [ɾunako], (b) [ɾunato], (c) [ɾuʃi̥ko], and (d) [ɾusu̥ko]. Frequency ranges: (a, b) 0–5000 Hz, (c, d) 0–8000 Hz.

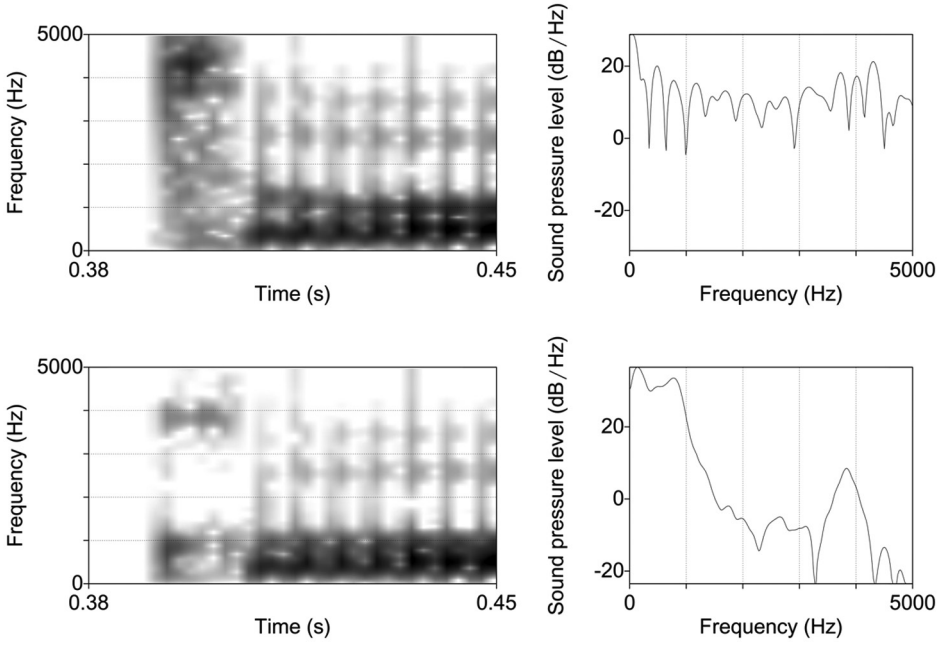

Figure 6. From bottom to top, spectrograms of the beginnings of the [ko] and [to] portions of the endpoints of the [k-t] continuum and spectra of 25 ms Gaussian windows centered on the stop bursts.

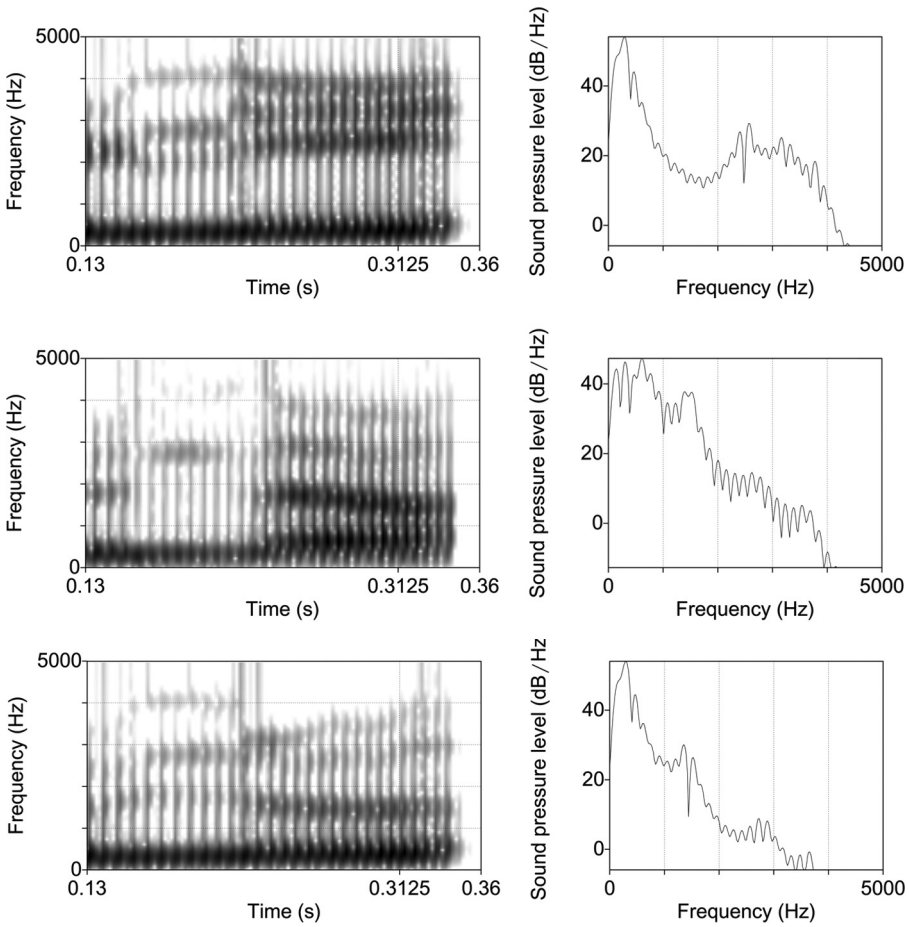

Figure 7. From bottom to top, spectrograms of the [ni, na, nu] portions of the voiced vowel stimuli and spectra from 25 ms Gaussian windows centered at 0.3125 ms (late in the vowels).

18 years, and they reported no hearing or speaking disorders nor any significant exposure to any language other than Japanese and English, respectively, before beginning school.

3.1.5 Listening conditions. The Japanese participants were run in a quiet room, while the English participants were run in a sound-attenuated room. All stimuli were output at 16 kHz and presented binaurally at a comfortable volume, via Beyerdynamic DT 250 80 Ohm headphones to the Japanese listeners and via Sennheiser HD 280 64 Ohm headphones to the English listeners. Cedrus SuperLab Pro software (version 2.0.4) presented all stimuli and cues and logged all responses. All the participants used Cedrus RB-834 response boxes to enter their responses.

3.1.6 Trial presentation structure. Each test trial began with a single stimulus. When it finished playing, two color-coded, language-specific visual prompts appeared onscreen, to cue listeners to respond: *katakana* representations of "ko" and "to" for the Japanese listeners, and the letters "k" and "t" for English listeners. The prompts were displayed until the listener responded or for 1500 ms, whichever was sooner. The inter-trial interval was 750 ms.

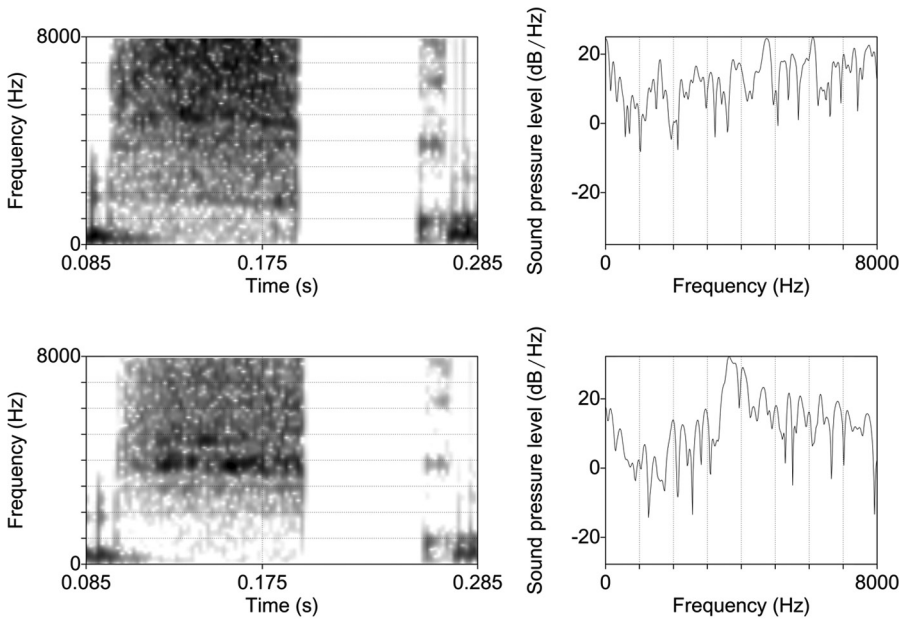

Figure 8. From bottom to top, spectrograms of the [ʃi̥] and [sʉ̥] portions of the fricative-voiceless vowel stimuli and spectra from 25 ms Gaussian windows centered at 0.175 ms (late in the fricative).

Listeners were first trained with 4 randomized repetitions of each continuum endpoint, with correct answer feedback displayed onscreen for 500 ms after their response. Once training ended, listeners were presented with seven blocks without feedback, each consisting of the entire stimulus array. A total of 21 repetitions were presented of the intermediate stimuli (steps 3–10), 14 of the near-endpoints (2, 11), and 7 of the endpoints (1, 12).

Participants took short breaks between blocks throughout the experiment, and every session was finished within 60 minutes.

3.1.7 Instructions. Listeners were told to pay attention to the final syllable, and to identify it as "ko" or "to" if Japanese or as beginning with "k" or "t" if English. Participants were instructed to respond as quickly as possible.

For half the participants, the "k(o)" response was assigned to the left button and the "t(o)" response to the right; their assignment was reversed for the other half. Although this variable was included in the statistical analyses, no results are reported for it.

3.2 Results

3.2.1 Japanese listeners. Figure 9 plots the mean proportion of "to" responses for each step along the [k-t] continuum in the three vocalic contexts for the Japanese listeners. The total proportions of "to" responses across the entire [k-t] continuum for each context are listed in Table 3—all statistics were carried out on these values. These proportions did not differ between the three vowels, $F < 1$. Figure 10 shows the mean proportions of "to" responses obtained from the Japanese listeners in the two fricative contexts. [ʃi̥] induced significantly more "t" responses than [sʉ̥], $F(1, 18) = 9.05$, $p < .01$.

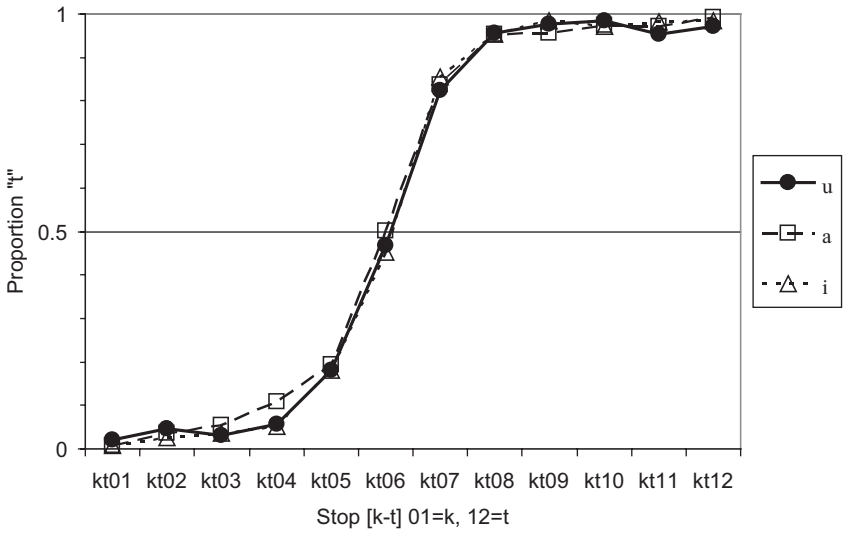

Figure 9. Mean proportions of Japanese listeners' "to" responses to each step along the [k-t] continuum in the three voiced vowel contexts.

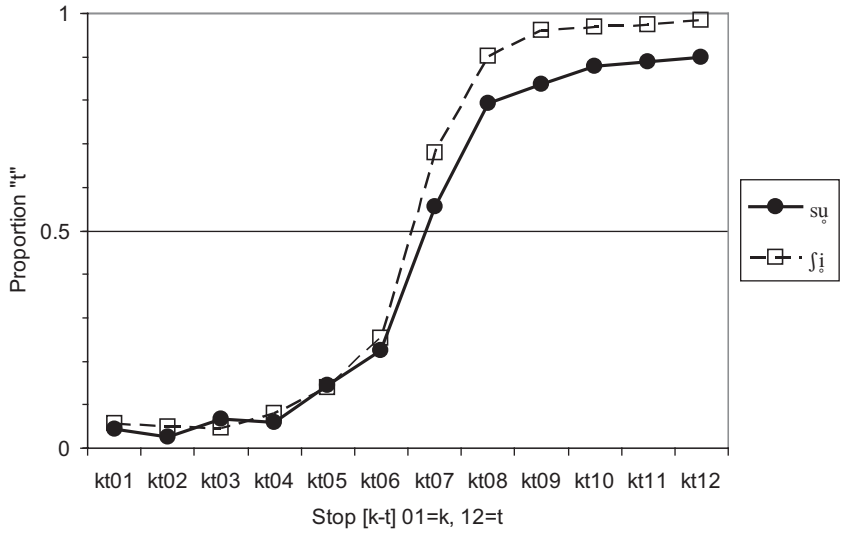

Figure 10. Mean proportions of Japanese listeners' "to" responses in the two fricative-devoiced vowel contexts, [ʃi̥] and [sụ̥].

3.2.2 English listeners. Figures 11 and 12 display the responses obtained from the English listeners in the same formats. They responded "t" most often after [u], less often after [a], and least often after [i] (Figure 11). The effect of vowel was significant, $F(2, 36) = 5.116$, $p = .011$. The difference between [u] and [i] was significant, $F(1, 18) = 15.375$, $p = .001$, but the difference between [u] and [a] was only marginally significant, $F(1, 18) = 3.935$, $p = .063$, and that between

Table 3. Mean total proportions (95% confidence intervals) of "t" responses across the [k-t] continuum as function of the preceding context and the listeners' native language, Experiment 2.

Context	Japanese	English
nu	0.539 (0.032)	0.595 (0.047)
na	0.549 (0.035)	0.548 (0.050)
ni	0.540 (0.035)	0.535 (0.030)
sụ	0.451 (0.068)	0.430 (0.070)
ʃi̥	0.508 (0.046)	0.476 (0.060)

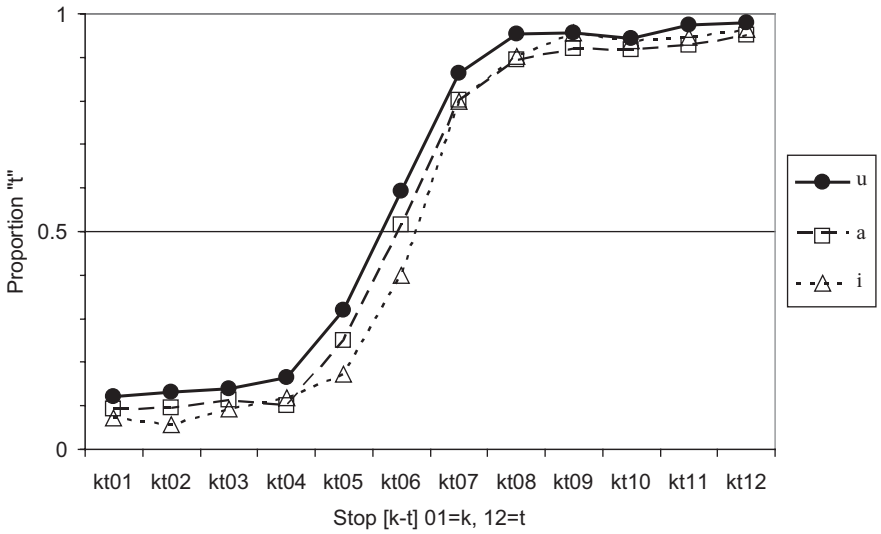

Figure 11. Mean proportions of English listeners' "t" responses to each step along the [k-t] continuum in the three voiced vowel contexts.

[a] and [i] did not even reach marginal significance, $F < 1$. English listeners also responded "t" significantly more often after [ʃi̥] than [sụ̥], Figure 12: $F(1, 18) = 6.798$, $p = .018$.

3.3 Summary and discussion

Preceding fricative-devoiced vowel intervals but not preceding voiced vowels influenced Japanese listeners' stop place judgments, while both fricative-devoiced vowel intervals and voiced vowels influenced the English listeners'. That a preceding [ʃi̥] caused listeners to respond "t" more often than a preceding [sụ̥] confirms the prediction of the auditory contrast account (Hypothesis 1) that listeners would hear a stop as contrasting with the spectrum of its fricative-devoiced vowel context.

These results also disconfirm Hypothesis 2, which predicts that listeners should respond "t" no more often after [ʃi̥] than [sụ̥] because they compensate for coarticulation with the devoiced vowel as well as the fricative, and the two perceptual adjustments would cancel one another out. Experiment 1 showed that stops do coarticulate with preceding voiceless vowels as well as with preceding fricatives. Because listeners have acoustic evidence that should lead them to compensate for

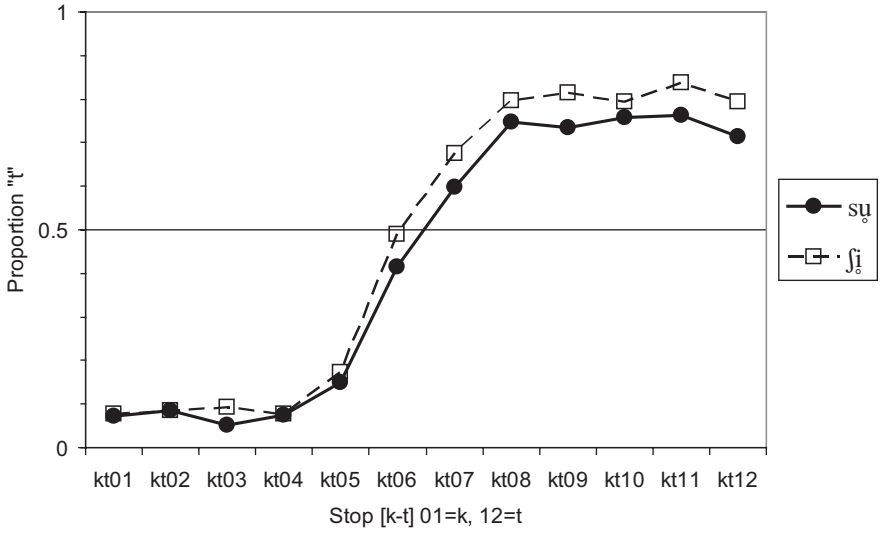

Figure 12. Mean proportions of English listeners' "t" responses in the two fricative-devoiced vowel contexts, [ʃị̥] and [sụ̥].

coarticulation with the voiceless vowels, it is surprising that they do not do so under the compensation account.

One might ask whether we should expect Japanese listeners **not** to compensate for coarticulation with a voiceless vowel, given that they also apparently do not compensate for coarticulation with preceding **voiced** vowels. This finding actually compounds the challenge to the compensation for coarticulation account rather than removing it. First of all, there is no competing articulation that they might instead compensate for when the vowel is voiced. Second, the English listeners' stop place biases did shift as a function of the preceding voiced vowel's quality, and if those shifts are to be attributed to compensation for coarticulation, Japanese listeners should compensate similarly. Third, Experiment 1 showed that the coarticulatory effects of voiced and voiceless vowels on following stops do not differ much in extent from one another nor are they appreciably less than those of fricatives, so we should expect listeners to compensate for them if they compensate for anything.

The absence of a voiced vowel effect on Japanese listeners still challenges the auditory contrast account, because the stops' spectra should contrast with the spectra of voiced [i] and [u] just as much for these listeners as the English listeners.

Whatever the lack of an effect of voiced vowels on Japanese listeners' stop percepts may mean for testing the competing hypotheses, we must still try to explain it. We try to do so in the next experiment.

4 Experiment 3: Varying the quality of the voiced vowel

In Experiment 2, Japanese listeners did not respond "t" more after [u] than after [i], unlike the English listeners. We consider two hypotheses here. [u] may not induce Japanese listeners to respond differently compared to [i] because the Japanese high back vowel does not concentrate energy particularly low in the spectrum (see Figure A.1 in the Appendix, but cf. Figure 7 bottom panel). Because English [u] or [ʊ] are the closest vowels to Japanese [u], English listeners may map

the Japanese vowel on to one of these vowels. They might then treat this vowel as though its spectrum concentrated energy as low as their own high back vowels do.

Alternatively, Japanese listeners may be insensitive to the spectral properties of [u] because it is the epenthetic vowel inserted to break up illicit consonant clusters in nonce words and loanwords. Dupoux et al. (1999) showed that Japanese listeners hallucinate this vowel in such clusters. Because the stimuli in Experiment 2 were nonce words, it would not be surprising if Japanese listeners were to treat [u] as epenthetic. English listeners would not treat [u] as epenthetic, nor would they distinguish it phonologically from any other full vowel nucleus.

These alternatives arise from different starting assumptions. The first attributes the Japanese listeners' indifference to the phonetics of Japanese [u], and the English listeners' sensitivity to the phonetic characteristics of the closest English vowels. The second attributes it to differences in this vowel's phonological status between the two languages.

These alternatives were tested in two ways. First, we added long [u] and [i] to the voiced vowel contexts. This manipulation tests the second alternative: a long vowel cannot be epenthetic, even if a short one can be. Second, we also added short and long mid vowels, [e, e:, o, o:]. This manipulation contrasts front and back—or spectrally high and low—vowels that are not epenthetic. Moreover, [o] has a lower F2 than [u] in Japanese (see Table A.1 in the Appendix), so this manipulation tests the phonetic explanation for why the difference between contextual voiced [u] and [i] was perceptually irrelevant to Japanese listeners.

4.1 Method

The stimuli had the form [ɾunVCo], where the V was one of [i, i:, u, u:, e, e:, o, o:], and the C was from the same 12-step [k-t] continuum used in Experiment 2. The short high vowels were the same as those used in Experiment 2. Parameter values used to synthesize the mid vowel stimuli were extracted from representative tokens of the mid vowels produced in [ɾunV$_i$hV$_i$] nonce words by the second author in the frame *dʒaa* ___ *de onegai* where V$_i$ was either [e] or [o]. Recordings were made using the same procedures as in Experiments 1 and 2. The parameter values were the same for both long and short vowels. The duration of the short vowels was 130 ms following the preceding [n]; the vowel's duration was increased to 250 ms to produce the long vowels.

These stimuli were presented to 17 Japanese listeners and 26 English listeners, who met the same criteria as participants in Experiment 2. None of them participated in Experiments 1 or 2. Both groups of listeners were recruited from the University of Massachusetts, Amherst community. They earned either course credit or $10/hour for participating.

The procedures were identical to those used in Experiment 2. Following endpoint training with feedback, listeners responded to 4 repetitions of each endpoint stimulus (1, 12), 8 repetitions of the near-endpoints (2, 11), and 12 repetitions of the intermediate stimuli (3–10).

4.2 Results

As in Experiment 2, statistical analyses were run on the total proportion of "t" responses across the continuum (Table 4). The independent variables were vowel height (high [i, i:, u, u:] versus mid [e, e:, o, o:]), backness (front [i, i:, e, e:] versus back [u, u:, o, o:]), and quantity (short [i, e, u, o] versus long [i:, u:, e:, o:]).

Figure 13 shows that Japanese listeners responded "t" more often after back than front vowels, although the difference in total proportions is modest (back 0.547 ± 0.019 versus front 0.526 ± 0.017). Although inspection of Table 4 shows that more "t" responses were obtained after the mid

Table 4. Mean total proportions (95% confidence intervals) of "t" responses across the [k-t] continuum as function of the preceding context and the listeners' native language, Experiment 3.

Context	Japanese	English
nu	0.527 (0.047)	0.582 (0.039)
nuu	0.530 (0.049)	0.546 (0.034)
no	0.566 (0.030)	0.533 (0.033)
noo	0.564 (0.029)	0.528 (0.035)
ni	0.500 (0.034)	0.487 (0.030)
nii	0.520 (0.043)	0.509 (0.032)
ne	0.538 (0.035)	0.454 (0.029)
nee	0.546 (0.031)	0.486 (0.030)

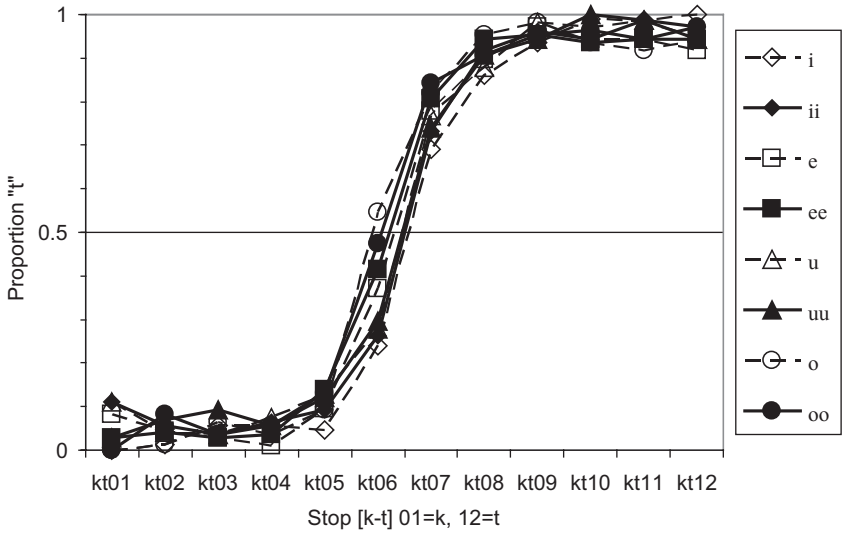

Figure 13. Mean proportion of "to" responses obtained from Japanese listeners in voiced vowel contexts [i, i:, e, e:, u, u:, o, o:].

vowels [o, o:, e, e:] than after the high vowels [u, u:, i, i:], only vowel backness was significant, $F(1, 15) = 13.945$, $p = .002$. Because the difference in response frequencies after mid versus high vowels is clearest for the intermediate stimuli, the analysis was rerun on the total proportion of "t" responses for steps 3–10 along the continuum. These are also the stimuli for which we have the greatest number of responses, so the estimates of "t" response proportions are most reliable for them. Backness was still significant, $F(1, 15) = 10.824$, $p = .005$, and height became marginally significant, $F(1, 15) = 4.441$, $p = 0.052$.

Figure 14 shows that English listeners responded "t" more often after the back than front vowels. They also responded "t" more often after long than short back vowels, but more often after short than long front vowels. Table 4 shows that these listeners also responded "t" more often after high than mid vowels, but the effects of quantity differ between vowel heights: listeners responded "t" more often after short than long mid vowels but differ very little in how often they respond "t" after short compared to long high vowels. The main effects of both backness and height were

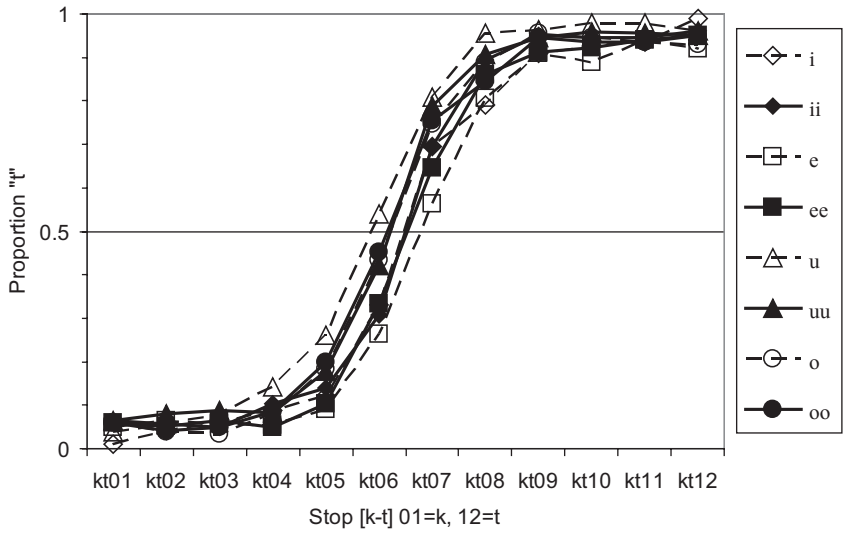

Figure 14. Mean proportion of "t" responses obtained from English listeners in voiced vowel contexts [i, i:, e, e:, u, u:, o, o:].

significant, backness: $F(1, 24) = 46.766$, $p < .001$; height: $F(1, 24) = 7.657$, $p = .011$. The backness by quantity interaction was clearly significant, $F(1, 24) = 30.006$, $p < .001$, but that between height and quantity was only marginally so, $F(1, 24) = 3.607$, $p = .07$. Reanalysis of total proportions of "t" responses just to steps 3–10 did not change the results for the English listeners.

4.3 Summary and discussion

Japanese listeners responded "t" more often after back than front vowels, and for the intermediate steps in the [k-t] continuum, more often after mid than high vowels. Their responses were, however, unaffected by the quantity of the preceding vowel. The absence of any quantity effect suggests that it is not the possibly epenthetic status of [u] that made Japanese listeners insensitive to the difference between [u] and [i] in Experiment 2. Instead, this result, together with the finding that [o] induced more "t" responses than [u] suggests that the phonetics of [u] were instead responsible. Because [u]'s F2 and F3 were both higher than [o]'s, auditory contrast would have produced a weaker bias toward "t" responses after [u] than [o]. If [u] is articulated farther front than [o], i.e. with its constriction at the back of the palate as compared to in the upper pharynx as it is in other languages (Wood, 1982), compensation for coarticulation would also predict a weaker bias after [u] than [o]. This evidence is not conclusive, however, as F2 and F3 were both higher at the end of [e] than [u] and [e] is certainly articulated farther front than [u], yet [e] induced a stronger bias toward "t" responses than [u] did.

The English listeners' responses amplified the pattern observed in Experiment 2 and still differed from the Japanese listeners' responses. Ordering the vowels by the frequency of "t" responses, [u] > [o] > [i] > [e], shows more "t" responses after back than front vowels and within these classes, more after high than mid vowels. The effect of vowel quantity was idiosyncratic: listeners responded "t" more often after short than long [u], but more often after long than short [e], and the number of "t" responses did not differ noticeably after long versus short [o] or [i].

It remains puzzling why the voiced vowel contexts affect Japanese listeners' judgments of the place of the following stops so little and so differently from how they affect English listeners' judgments. Although the English listeners responded to the same stimuli as the Japanese listeners, they apparently did not respond to the vowel's concrete phonetic characteristics but instead to their abstract, even idealized phonological cross-classification by the features [back] and [high].

The difference between the responses of the two groups should perhaps not be so surprising. The stimuli were modeled on actual Japanese utterances and as such would have fit Japanese listeners' phonetic expectations far better than English listeners'. This better fit would make it easier for Japanese listeners to respond to the familiar phonetic differences between the various vowel qualities, in particular [o]'s F2 being lower than [u]'s. The poorer fit to English listeners' expectations might have forced them to rely instead on their phonological analysis of these vowels. None of this is anything more than speculation at this point, but it suggests that future research should examine the effects of differences in phonetic experience on how contexts influence listeners' response biases (see also Norris, McQueen, & Cutler, 2003).

5 Summary and general discussion

5.1 Summary

Experiment 1 presented acoustic evidence that [k] and [t] coarticulate with preceding fricatives and with voiceless as well as voiced vowels in Japanese, and thus that there is something to compensate for when they are heard in these contexts.

Experiment 2 compared two hypotheses: (1) English and Japanese listeners would both respond "t" more often after [ʃi̥] than [su̥], because their percept of the stops' spectra would contrast auditorily with the fricatives' spectra versus (2) Japanese listeners would not respond "t" more often after [ʃi̥] than [su̥] because compensating for coarticulation with the fricative would cancel out compensating for coarticulation with the devoiced vowel. The results supported the first hypothesis. Yet the Japanese and English listeners differed in the effects of preceding voiced vowels differing in place: Japanese listeners' stop place judgments were unaffected by differences between these vowels, while English listeners responded "t" more often after [u] than [a] and more often after [a] than [i]. The English listeners' response biases after these voiced vowels support both hypotheses, while the Japanese listeners' lack of biases in these contexts support neither.

Experiment 3 sought to determine why Japanese listeners' judgments were not affected by differences between preceding voiced vowels in Experiment 2. It tested two further hypotheses: (3) [u] did not induce more "t" responses than [i] because Japanese listeners treat [u] as an epenthetic vowel, versus (4) Japanese [u] does not differ enough acoustically from [i] to produce sufficient perceptual contrast between the vowels' and the following stops' spectra. The two hypotheses were tested by adding long and mid vowels as voiced vowel contexts. Unlike short [u], long [u:] could not be interpreted as epenthetic. The mid back vowel [o] has a lower F2 and F3 than [u] in Japanese, so it might serve a phonetically better foil to [i] than [u]. On the one hand, as predicted by the fourth hypothesis, Japanese listeners did respond "t" more often after [o] than [i]. On the other hand, contrary to the third hypothesis, they did not respond differently after long than short [u]. English listeners responded "t" more often after both back vowels than both front vowels, and within the back and front classes, more often after the high than the mid vowel. This difference suggests that the English listeners were responding in terms of the phonological categories to which they assigned these vowels rather than the phonetic characteristics of the vowels that drove the Japanese listeners' responses. The next two sections compare the successes of the first two hypotheses in more detail.

5.2 Compensation for coarticulation?

The effects of [ʃ] versus [s] on categorization of a following [k-t] continuum had originally been explained by Mann and Repp (1981) as compensation for coarticulation: listeners would identify a stop intermediate between [k] and [t] as "t" more often after [ʃ] than [s] because they undo the perceived backing of the stop by [ʃ]. According to Hypothesis 2, listeners in Experiment 2 had two sources of coarticulation to compensate for, which would cancel one another out, because [ʃ] in Japanese retains acoustic vestiges of devoiced [i̥] and [s] retains vestiges of devoiced [u̥]. The Japanese listeners' response biases following [ʃi̥] and [su̥] in Experiment 2 disconfirmed this prediction. Instead of cancellation, they responded "t" more often after [ʃi̥] and [su̥].

An alternative version of the compensation hypothesis predicts that listeners would compensate for coarticulation with the fricatives alone. For this alternative to be viable, listeners would have to fail to perceive the devoiced vowels during the fricatives and their coarticulatory perturbation of the stop's acoustics. Such a failure is perhaps less surprising for English listeners, because English phonotactics does not require a vowel to separate heterorganic consonants, and /ʃ/ and /s/ are distinct phonemes in English rather than being allophones conditioned by /i/ and /u/, respectively. They would therefore have no reason to expect vowels to occur in what sound like [ʃk, ʃt, sk, st] clusters, and might miss the perhaps subtle phonetic evidence of their presence. But precisely because these are characteristics of Japanese phonotactics and allophony, it **is** surprising that Japanese listeners would not notice the phonetic evidence for the devoiced vowels and compensate for their expected coarticulatory effects. To account for the Japanese listeners' responses, the compensation account would have to be amended *ad hoc* to permit listeners to ignore an articulation that produces measurable acoustic consequences and coarticulatory effects.

That failure of compensation for coarticulation account raises the following questions: (1) is the devoiced vowel's articulation present and perceptible during the fricative's pronunciation, (2) do following stops coarticulate with these devoiced vowels, and (3) is the perceptible effect of coarticulation with the vowel comparable in extent to that of coarticulation with the fricative?

Nakamura's (2003) electropalatographic evidence shows that the fricative and stop gestures in a sequence where the intervening vowel devoices, for example, in [su̥k], overlap, although the extent of overlap varies considerably. Nakamura nonetheless finds traces of the devoiced vowels' oral gestures in his data. Beckman and Shoji (1984) and Tsuchida (1994) show that these vowels' oral gestures are perceptible.

Experiment 1 showed that [k, t] do coarticulate with preceding voiceless vowels, as well as with the preceding fricative. Tsuchida (1994) found that a fricative coarticulates with the next consonant when the intervening vowel devoices, but she did not determine whether the stop coarticulates with the devoiced vowel nor if it does whether that coarticulation changes the stop's acoustics differently from coarticulation with the fricative. If the acoustic effects of coarticulating with the two segments are opposite in direction but equal in size, as we showed in Experiment 1, then we would expect them to cancel out perceptually and listeners' judgments of a following [k-t] continuum not to be biased by these contexts.

No study compares the perceptibility of coarticulation with the preceding fricative versus the devoiced vowel, other than our own Experiment 2, which showed that neither Japanese nor English listeners' responses were influenced by the devoiced vowels' backness.

What remains is the possibility that stops only coarticulate perceptibly with the preceding fricative, in which case the compensation for coarticulation account makes exactly the same predictions as the auditory contrast account, and we are back where we began.

5.3 Auditory contrast?

Is Hypothesis 1 the only one left standing? Its predictions were confirmed by both Japanese and English listeners' response biases following [ʃi̥] and [su̥], and by English listeners' response biases following voiced vowels, but they were disconfirmed by Japanese listeners' failure to show any response biases following voiced vowels in Experiment 2. That failure may undermine the compensation account, too, but it still needs to be explained.

The results of Experiment 3 suggested a tentative explanation, namely, that energy is concentrated too high in the spectrum of Japanese [u] for it to induce perceptibly different contrast than [i]. This explanation is supported by the finding that Japanese listeners responded "t" less often after [u] than [o], whose F2 and F3 were lower than the [u]'s. English listeners, however, responded "t" more often after [u] than [o]. Japanese listeners might have been more sensitive to this phonetic difference between [u] and [o] because the stimuli's formant frequency values were taken from Japanese models and should thus be more familiar to Japanese listeners (see Norris, McQueen, & Cutler, 2003, for evidence that even short-term changes in listeners' experience can change their response biases). English listeners might instead have assimilated Japanese [u] to the English category, /u/, that has the lowest energy concentration. Even if these speculative explanations hold up, they clearly go far beyond the explanation that relies on auditory contrast between target and context alone. For other challenges to the contrast account, see Fowler, Brown, and Mann (2000), Viswanathan, Magnuson, and Fowler (2008, 2010), and Viswanathan, Fowler, and Magnuson (2009).

5.4 Concluding remarks

The current experiments tested competing accounts of why a speech sound is perceived differently from its context: the auditory contrast account and the compensation for coarticulation account. Their results do not support the most general version of the compensation account because listeners did not compensate for coarticulation with the devoiced vowels in the [ʃi̥] and [su̥] contexts in categorizing the following stops, even though the devoiced vowels would have been detectable and following stops do coarticulate with them. Both Japanese and English listeners' response biases in these contexts support the contrast account instead, as do English listeners' response biases in the voiced vowel contexts. Neither the contrast nor the compensation account is supported by the absence of any difference in Japanese listeners' response biases across the voiced vowel contexts in Experiment 2. Both accounts are supported by the presence of different response biases across voiced vowel contexts in Experiment 3. The results of that experiment indicate, however, that neither auditory contrast nor compensation for coarticulation between target and context fully explains the perceptual effects of the target sound's context on its perception; differences in linguistic experience appear to modulate how much the percept of the stop's spectrum or place of articulation depends on the spectrum or place of articulation of the preceding voiced vowel.

Acknowledgements

Experiment 2 was presented at the Providence meeting of the Acoustical Society of America (Mash & Kawahara, 2006), a joint meeting of the University of Massachusetts Phonology Group and the MIT Phonology Circle on 29 March 2008, and at McGill University on 8 October 2008. We are grateful to the audiences for their very helpful comments. Our colleagues in the Phonetics Laboratory, Eve Brenner-Alsop, Richard Cruz-Yi, Michael Key, Anne Pycha, and Sarah Watsky, have provided the most sustained and helpful commentary and reaction. The reactions of Anne Cutler and two anonymous reviewers to earlier versions of this article have also been of great help to us in preparing the current version. We retain all rights to any

remaining errors. This research was supported by NIH grant R01-DC006241 to the first author, which is gratefully acknowledged.

Notes

1 [ʃu̥] is the surface realization of /sju/ before a voiceless consonant.
2 Speaker was included to soak up variance; differences between speakers were otherwise not of particular interest and will not be discussed. Stop place always had a significant effect on the dependent measures, because all three measures were all always substantially higher for [t] than [k]. Therefore, stop place is discussed only when it interacted significantly with the contextual variables.
3 This difference in the duration of the silent interval was unintentional, but should exaggerate any perceptual adjustment listeners make for voiced as compared to voiceless vowel contexts.

References

Amano, S., & Kondo, T. (2000). *NTT database series: Lexical properties of Japanese* (2nd release). Tokyo: Sanseido.

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

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

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*, 1001–1017.

Boersma, P., & Weenink, D. (2007). Praat: Doing phonetics by computer (Version 4.5.16) [Computer program]. Retrieved February 18, 2007, from http://www.praat.org/

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*(6), 149–179.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578.

Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*, *16*(5–6), 491–505.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception and Psychophysics*, *68*, 161–177.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 877–888.

Han, M. (1962). Unvoicing of vowels in Japanese. *Onsei on Kenkyuu* [Studies in Phonetics], *10*, 81–100.

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

Haraguchi, S. (1977). *The tone pattern of Japanese: An autosegmental theory of tonology*. Tokyo: Kaitakusha.

Hattori, S., Yamamoto, K., Kohashi, Y., & Fujimura, O. (1957). Nihongo no boin [Japanese vowels]. *Bulletin of the Kobayasi Institute of Physical Research*, *7*, 69–79.

Jakobson, R., Fant, C. G. M., & Halle, M. (1951). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *108*, 1252–1263.

Keating, P. A., & Huffman, M. K. (1984). Vowel variation in Japanese. *Phonetica*, *41*(4), 191–207.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception and Psychophysics*, *68*, 178–183.

Lotto, A. J., & Kluender, K. R. (1998). Gestural contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*, 602–619.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on the perception of the [ʃ]-[s] distinction. *Perception and Psychophysics*, *28*, 213–228.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, *69*, 548–558.

Mash, D., & Kawahara, S. (2006). Sequential contrast versus compensation for coarticulation in English and Japanese. *Journal of the Acoustical Society of America*, *119*, 3423 (Abstract).

McCawley, J. D. (1977). Accent in Japanese. In L. M. Hyman (Ed.), *Studies in stress and accent* (pp. 251–302). Los Angeles: University of Southern California Occasional Papers in Linguistics.

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

Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., & Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *Journal of the Acoustical Society of America*, *124*, 576–588.

Nittrouer, S., Studdert-Kennedy, K., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech, Language, and Hearing Research*, *32*, 120–132.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *64*, 1358–1368.

Tsuchida, A. (1994). Fricative-vowel coarticulation in Japanese devoiced syllables: Acoustic and perceptual evidence. *Working Papers of the Cornell Phonetics Laboratory*, *10*, 145–165.

Tsuchida, A. (1997). *Phonetics and phonology of Japanese vowel devoicing*, PhD dissertation, Cornell University.

Vance, T. (2005). *The sounds of Japanese*. Cambridge: Cambridge University Press.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin and Review*, *16*, 74–79.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2008). Compensation for coarticulation may reflect gestural perception: Evidence from a critical examination of the effects of non-speech contexts on speech categorization. *Eleventh Conference on Laboratory Phonology* (pp. 147–148). Wellington, New Zealand (Abstract).

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005–1015.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the /s-ʃ/ boundary. *Journal of the Acoustical Society of America*, *69*, 275–282.

Wood, S. (1982). *X-ray and model studies of vowel articulation*, PhD dissertation, Lund University.

Appendix

Table A.1 lists the average values for F2 in Japanese vowels reported in three studies. The high back vowel in Japanese is not rounded, but the mouth opening is compressed vertically, and as a result, this vowel differs in quality from high back unrounded [ɯ] (Vance, 2005, Chapter 3). All three studies, as well as our own (Figure A.1) found that F2 was as high in this vowel as in [a].

Table A.1. Average F2 values for the Japanese vowels reported by Hattori, Yamamoto, Kohashi, & Fujimura (1957), Keating & Huffman (1984), and Nishi, Strange, Akahane-Yamada, Kubo, & Trent-Brown (2008).

Vowel	Hattori et al.		Keating & Huffman	Nishi et al.
	Male	Female		
i	2300	2930	1954	2007
u	1180	1430	1419	1171
e	not measured		1720	1785
o	not measured		1136	805
a	1180	1450	1383	1182

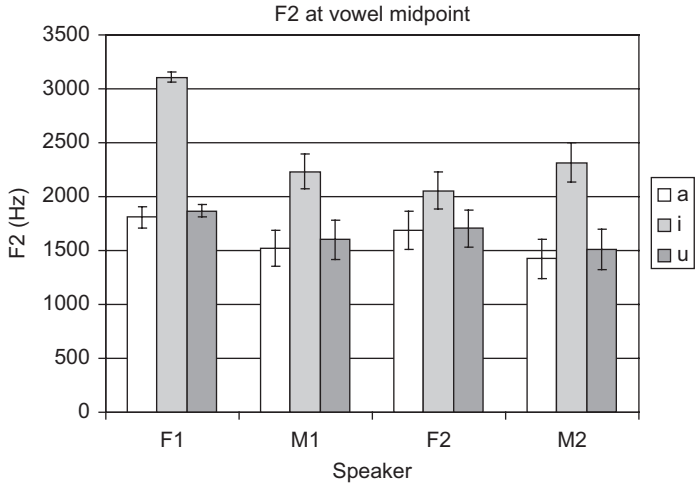

Figure A.1. Mean F2 values (95% confidence intervals) at vowel midpoint for [a, i, u] produced by two male (M1, M2) and two female (F1, F2) speakers of Tokyo Japanese (see Experiment 1 for a description of the materials in which these vowels were pronounced).