Auditory contrast and autonomy in context effects: Data from Japanese and English listeners[?]

John Kingston, Shigeto Kawahara, Daniel Mash, and Della Chambless

Linguistics Department

University of Massachusetts

150 Hicks Way, 226 South College

Amherst, M A 01003-9274

jkingston@linguist.umass.edu

1-413-545-6833, fax -2792

[?] The results reported in this paper were first presented in a poster at the Providence meeting of the Acoustical Society of America (Mash and Kawahara 2006). The research reported here was supported by NIH grant R01-DC006241 to the first author.

Abstract

The three experiments reported here test two hypotheses. First, auditory contrast between a sound and and its context alters the sound’s perception rather than the listener compensating for coarticulation with the sound’s context or using his phonological knowledge to infer a perceptually relevant property of the context. Second, auditory contrast arises during a processing stage that is autonomous from a stage when listeners apply their linguistic knowledge to evaluating the signal. These hypotheses are tested by measuring Japanese and English listeners’ response biases when categorizing a [k-t] continuum following [s] and [?]. In Japanese, these fricatives are allophones of /s/ conditioned by the voiceless vowels [u?] and [i?], respectively. If auditory contrast with the fricatives’ spectra biases listeners’ stop place percepts, then both Japanese and English listeners should respond “k” more often after [s] than [?]. If they instead compensate for coarticulation with the conditioning vowels, then they will either respond “k” more often after [?] instead, or compensating for coarticulation with the voiceless vowel will cancel out compensation for coarticulation with the fricative. Listeners from both languages responded “k” more often after [s] than [?]. This result not only rules out the compensation for coarticulation but also Japanese listeners’ inferring that the voiceless vowels are present from the fricatives’ pronunciations, and treating them as the perceptually relevant context. The second and third experiments tested the effects of varying the syllabification of the stop with respect to the preceding context and revealed that neither Japanese nor English listeners' stop place biases depend on syllabification. This result suggests that the stops were not syllabified with their contexts at the stage when auditory contrast with the contexts altered the percepts of stop place, and thus provides further support for the second hypothesis.

Key words: Auditory contrast, compensation for coarticulation, phonological inference

The identification of speech sounds can be influenced by neighboring sounds. For example, Mann & Repp (1981) show that English listeners identify a stop that is ambiguous between [t] and [k] more often as “t” after [?] than after [s]. Perceptual shifts caused by adjacent sounds have been found across many classes of targets and contexts, and in all published cases the target is identified as differing from its context (Lotto & Holt, 2006).

At least two theories can explain response shifts such as that caused by the contextual fricative in Mann & Repp (1981). Mann & Repp themselves propose that listeners expect a stop articulated after the palatoalveolar [?] to be pronounced further back than after the alveolar fricative [s], perceptually undo the effects of this backing on the stop’s acoustics, and hear the stop as “t”. This perceptual adjustment is referred to as “compensation for coarticulation” (see also Fowler, 2006). Alternatively, the stop may be perceived as differing *auditorily* rather than articulatorily from it context. In the example, a stop ambiguous between [t] and [k] literally sounds higher in frequency—hence more like [t]—after the relatively low spectral concentration of energy in the fricative [?]. This perceptual adjustment is called “auditory contrast”. Compensation for coarticulation relies on the listener perceiving (and perhaps knowing) that the acoustics of the stop are intermediate in value between [t] and [k] because the stop is coarticulated with the fricative. In compensating for this coarticulatory effect, the listener judges the stop as differing in backness from its context. Sequential spectral contrast automatically exaggerates the perceived difference between the acoustics of the stop and the fricative during auditory processing, and the listener need not perceive nor know anything about how or even if those sounds coarticulate with one another.

Though both theories explain the results of Mann & Repp (1981) equally well, auditory contrast alone has the additional virtue of being able to explain the comparable perceptual shifts obtained when the contexts are not speech sounds (Lotto & Kluender, 1998; Holt, Lotto & Kluender, 2000) or when the listeners are not linguistically informed humans (Lotto, Kluender & Holt, 1997, cf. Fowler, Brown, & Mann, 2000; Fowler, 2006). A sound can contrast auditorily with a non-speech context as easily as a speech context, and such contrasts should arise as easily in the responses of the auditory systems of non-human as human listeners. However, human listeners would not perceive a speech sound as coarticulating with a non-speech context, nor would non-human listeners have any experience or knowledge of how one speech sound coarticulates with another. The stimuli used in the experiments reported in this paper provide another means of empirically distinguishing the predictions of the compensation and contrast accounts of the perceptual effects of context.

A second purpose of this paper is to determine when, if at all, during the course of speech perception the influence of context operates. The model proposed in Kingston (2005) distinguishes an initial auditory level of processing, which knows nothing about the listener's language, from a later stage when linguistic knowledge influences the listener's categorization of segments in the signal. The model implies, moreover, that the results of later, linguistically informed categorization do not feed back on the initial auditory processing. Auditory processing outputs perceived values along continuous auditory scales that represent the auditory transformations of the signal's acoustics. This model closely resembles Norris, McQueen, & Cutler's (2000) MERGE model, which also separates the initial auditory evaluation of the signal from the knowledge-driven interpretation of that evaluation, and likewise prohibits feedback from linguistic interpretation to auditory percept. In this respect, the autonomous model proposed here differs from interactive models of speech perception such as TRACE (McClelland & Elman, 1986; Elman & McClelland, 1988) in which linguistic knowledge influences the evaluation of the signal from the outset of processing. To the extent that the two-stage, autonomous model is correct, we can ask whether the target sound contrasts with its context during the earlier, linguistically naive auditory stage, as supposed above, or during the later linguistically informed stage.

It is possible to determine when contrast arises and whether the contrast or compensation account is correct using Japanese stimuli and listeners. The Japanese sibilant fricative is pronounced as [s] before the vowel [u],[1] but as [?] before [i]. Furthermore, the two high vowels devoice between two voiceless consonants, creating such apparent surface clusters as [sk], [st], [?k], and [?t] (Beckman, 1982; Han, 1962; McCawley, 1977; Beckman & Shoji, 1984; Tsuchida, 1994, 1997). However, Japanese phonology requires that a vowel intervene between any two non-homorganic consonants (Han, 1962; Itô, 1986, 1989). Japanese speakers apparently even hallucinate a vowel between the consonants of non-homorganic clusters when no acoustic vestige of a vowel occurs between them (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Dupoux, Pallier,

---

[1] We use the symbol [u] to represent this language's high back vowel. Though this vowel is not rounded, the mouth opening is compressed vertically, and as a result, F2 is low. This makes this vowel different in quality from that represented by the symbol for the high back unrounded vowel [?] (Vance, 2005, Chapter 3). Average F2 and F3 for four male and four female speakers of Japanese vowels reported in Hattori, Yamamoto, Kohashi & Fujimura (1957) are:

Vowel	Males		Female	
	F2	F3	F2	F3
[u]	1180	2500	1430	3100
[i]	2300	3150	2930	3550
[a]	1180	2750	1450	unmeasurable

Kakehi, & Mehler, 2001). Therefore, apparent phonetic clusters like [st] and [?t] must be represented phonologically as /sut/ and /?t/, respectively:

Phonological:	/sut/	/?it/
Phonetic:	apparent st = [su?]	apparent ?t = [?i?]

Beckman & Shoji (1984) argue that the vowel is devoiced in such sequences because speakers fail to adduct the glottis between two consonants which are themselves pronounced with the glottis abducted, but that speakers still produce the vowel's oral articulation. Nakamura (2003) also finds traces of the oral articulation of the devoiced vowels in his EPG data. The oral articulation of devoiced vowels is evident in these sequences primarily in the place of articulation of the strident fricative, which is alveolar when the vowel is [u?vs palatoalveolar when it is [i?, but secondarily in the fricative's greater duration (Han, 1994; cf. Beckman 1982). It is often even possible to see in a spectrogram the higher formants produced by the vowel articulation during the interval of fricative noise (Han, 1962; see also Figure 1 below). These characteristics of Japanese provide a means first of distinguishing empirically between the predictions of the contrast and compensation accounts of the perceptual effects of context, and second of determining whether sequential contrast arises during the early linguistically naïve auditory processing stage, as hypothesized, or the later linguistically informed stage.

We entertain here three competing hypotheses concerning listeners' judgments of the place of articulation of stimuli drawn from a [k-t] continuum following [su?] and [?i?]: (1) the percepts of stops' spectra contrast auditorily with the preceding fricatives' spectra, (2) listeners compensate for coarticulation with the voiceless vowel, the fricative, or both gestures, and (3) the percepts of the stops' spectra contrast with the spectra of vowels that listeners infer are present from their phonological knowledge of Japanese. These hypotheses and their predictions are laid out below.

Hypothesis 1: Auditory contrast.

If the stop's spectrum contrasts auditorily with its context's spectrum and this contrast arises during a linguistically naive level of processing, then only the predominant spectral properties of the context—the energy concentration in the fricative—would determine both Japanese and English listeners' biases in identifying the following stop (1 in Table I).

There are three versions of Hypothesis 2, one in which listeners compensate only for coarticulation with the voiceless vowel (Hypothesis 2a), a second in which they compensate for coarticulation with both the voiceless vowel and the fricative (Hypothesis 2b), and a third where they only compensate for coarticulation with the fricative (Hypothesis 2c). The first two alternatives, especially Hypothesis 2a, make different predictions about the effects of context on stop place judgments than Hypothesis 1, while the third makes the same predictions.

> Hypothesis 2a: Compensation for coarticulation with the voiceless vowel.
>
> If the devoiced vowel's oral articulation is present and perceptible in the signal simultaneously with the fricative, and listeners compensate for coarticulation with it alone, then both Japanese and English listeners' stop place judgments should be biased in exactly the opposite way from what Hypothesis 1 predicts (compare 2a with 1 in Table I).

Auditory contrast predicts more "t" responses after [?] than [s] from both Japanese and English listeners, while compensation for coarticulation with the voiceless vowel alone instead predicts more "t" responses after [u?than [i?from both groups of listeners. That is, both hypotheses predict that Japanese and English listeners will respond in the same way, but their predictions about how the two groups of listeners will respond are opposite to one another.

Because the fricative is more or less simultaneous with the voiceless vowel and the stop may coarticulate with it as well, we must consider an alternative to Hypothesis 2a in which listeners compensate for coarticulation with both sounds:

> Hypothesis 2b: Compensation for coarticulation with the voiceless vowel and the fricative.
>
> If listeners compensate for coarticulation with the fricative as well as the voiceless vowel, the two biases should cancel one another out and there would be no effect of context (combine 1 and 2a in Table I).

The compensation effects cancel one another out because the voiceless vowel that would pull the stop articulation forward, [i?], cooccurs with the fricative that would pull the stop back, [?], and vice versa. If coarticulation with one of these segments is perceived to be greater than that with the other, compensating for the weaker would not entirely cancel out compensating for the stronger, but the latter compensation should still be less than if listeners were not compensating for a competing coarticulation.

If the compensatory adjustments largely or completely cancel one another out, then we end up comparing a hypothesis that predicts no effect of context (2b) with one that predicts a positive effect (1) rather than two hypotheses that both predict positive if opposite effects (2a vs 1). As the absence of an effect cannot by itself confirm a hypothesis, it is necessary to add a condition to our experiment that would test a positive prediction of Hypothesis 2b. Fortunately, this is easy to do because this hypothesis predicts that listeners will still compensate for coarticulation with a preceding vowel in the absence of a competing fricative. That is, Hypothesis 2b predicts that listeners will compensate for coarticulation with voiced as well as voiceless vowels, and because no competing gesture is produced at the same time as the voiced vowel gesture, compensation for coarticulation with the voiced vowel gesture will not be cancelled. As a result, this hypothesis predicts that listeners will respond “t” more often after voiced [u] than voiced [i] because they will compensate for the expected backing of the stop’s articulation by the back vowel. Note that the auditorycontrast account also predicts more “t” responses after [u] than [i] but not because [u] is back but instead because its spectrum concentrates energy at lower frequencies than [i]’s does. Contrast with the vowel’s spectrum will cause a stop whose spectrum is intermediate between high [t] and low [k] to sound higher and more like [t] than [k].

Finally, listeners could compensate for coarticulation in a third way:

> Hypothesis 2c: Compensation for coarticulation with the fricative alone.
>
> If listeners compensate for coarticulation with the fricative, their stop place biases would be exactly the same as those predicted by Hypothesis 1, even though the mechanism that produces the bias is entirely different (predictions the same as 1 in Table I).

According to this hypothesis, both Japanese and English listeners would compensate for the coarticulation with the preceding fricative rather than the devoiced vowel, even if that devoiced vowel were pronounced, and following stops coarticulated with it. Because the compensation account makes the same prediction in this case as the contrast account, thjs is the least interesting of the three alternatives. Although it is premature to choose among these alternatives, we should note that only Hypothesis 2b follows directly from the compensation for coarticulation account if both the vowel and fricative gestures are produced and a following stop coarticulates with both of them. The other two alternatives, 2a and 2c, both make the as yet unmotivated assumption that listeners would ignore coarticulation with a gesture even if it is present in the signal, perceptible to the listener, and coarticulates with the stop.

The auditory contrast hypothesis’s predictions also differ from those of a third hypothesis (3 in Table I):

Hypothesis 3: Phonological contrast

> If Japanese listeners use their phonological knowledge to *infer* the presence of a particular vowel quality from the pronunciation of the fricative,[2] that vowel quality and not the fricative would bias their stop place judgment because it is the inferred vowel and not the fricative that is phonologically adjacent to the stop. English listeners have no phonological knowledge that would lead to such an inference, so the fricative should instead bias their stop place responses.

The third hypothesis derives from the assumption that listeners' percepts are determined at least in part by linguistically informed expectations about what should occur at a particular moment in the signal (McClelland & Elman, 1986; Elman & McClelland,

---

[2] [?] also occurs before the sequence [ju], and the [u] will devoice if it is followed in turn by a voiceless obstruent. Thus, the string [?t] is phonologically ambiguous: it may be the pronunciation of either /sit/ or /sjut/. However, Beckman & Shoji (1984) and Tsuchida (1994) have shown that Japanese listeners can distinguish between [??(< /si/) and [?u?(< /sju/) above chance level, which indicates that they can actually hear the devoiced vowel during the fricative. [st] is entirely unambiguous; it can only be the pronunciation of /sut/.

1988; and more specifically for Japanese, Dupoux et al., 1999, 2001). In particular, Elman & McClelland (1988) showed that listeners' "k-t" judgments were biased by a preceding ambiguous fricative whose identity as [s] or [?] could be predicted lexically (e.g. as [s] in the context *Christma_* but as [?] in the context *Spani_*). This finding raises the possibility that from the pronunciation of the strident fricative as [s] or [?], a Japanese listener knows the quality of the devoiced vowel between that fricative and a following voiceless stop, [u?] and [i ?], respectively, and uses that vowel context in identifying the stop. The first compensation for coarticulation alternative, Hypothesis 2a, and phonological contrast thus predict the same responses from Japanese listeners—more "t" responses after [s] than [?] because the relevant context is actually not the fricative but instead the devoiced vowel [u?]—but predict different responses from the English listeners because they have no phonological knowledge from which they could infer any vowel between the strident fricative and the following stop. In the phonological contrast hypothesis, there need be no perceptual vestige of the devoiced vowel in the signal, as the pronunciation of the strident fricative is sufficient to determine what its quality was.[3]

---

[3] The mechanism by which the inferred vowel quality biases stop place judgments could be auditory contrast or compensation for coarticulation, depending on whether the listener infers an auditory or an articulatory object. We see no reason why a vowel that is inferred rather than perceived must be one kind of object or the other, and in any case, the predictions are the same, more "t" responses after inferred [u ?] than [i ?].

The predictions of these hypotheses are summarized in Table I:

*** Please put Table I about here. ***

The predictions of auditory contrast hypothesis (1) differ from those of two alternative versions of the compensation for coarticulation hypothesis, regardless of whether listeners compensate for coarticulation with the voiceless vowel alone (2a) or the fricative as well (2b). They also differ from those of the phonological contrast hypothesis (3) for Japanese but not English listeners. The only version of the compensation for coarticulation account does not differ in its predictions from the auditory contrast account is alternative 2c, where listeners compensate for the fricative alone.

Testing the predictions of Hypotheses 1 vs 2a or 2b will answer the first question posed by this study: are context effects the product of auditory contrast between adjacent sounds or compensation for perceived coarticulation between them? If the quality of the devoiced vowel does not determine responses, we will be unable to distinguish between the predictions of Hypotheses 1 and 2c. Testing the predictions of Hypotheses 1 vs 3 will answer the second question: is the initial processing of speech sounds determined by their auditory qualities alone and not by any linguistic knowledge the listener might bring to bear?

We report first the results of an experiment designed to determine which of these predictions is correct. This experiment compared the proportion of “t” responses given by Japanese and English listeners to the members of a [k-t] continuum in two kinds of contexts: (1) following [su?] (< /su/) and [?i?] (</si/) and (2) following the voiced vowels [i, u, a]. Both Japanese and English listeners responded “t” more often after [?i?] than [su?], but only the English listeners’ response biases were affected by preceding voiced vowels, where they responded “t” more often after spectrally low voiced [u] than spectrally high voiced [i].

A surprising result of the first experiment was that Japanese listeners’ stop place judgments were not biased by a preceding voiced vowel. The second experiment tested the hypothesis that a voiced vowel had no effect on Japanese listeners’ judgments of the following stop’s place in the first experiment because Japanese phonotactics separates a vowel and a following singleton consonant into distinct syllables. Perhaps, neither auditory contrast nor compensation for coarticulation operates between adjacent sounds when they are not affiliated prosodically with one another. This experiment tested this hypothesis by lengthening the stop closure enough to cause Japanese listeners to hear the stop as a geminate rather than a singleton. Because geminates are obligatorily syllabified with preceding as well as following vowels in Japanese (Itô, 1986, 1989), this manipulation should have permitted a preceding voiced vowel to influence the Japanese listeners' percept of the stop's place if that influence depends on it being in the same syllable as the vowel for these listeners. The results showed clearly that the vowel did not affect Japanese listeners' stop place judgments any more when the stop closure was long enough to be a geminate than when it was short enough to be a singleton, disconfirming the hypothesis that the percept of a sound is only affected by a context that is prosodically affiliated with it.

Mann & Repp (1981) had shown that the affiliation of the context with the target sound can nonetheless affect its perception (see also Samuel & Pitt 2003 for a similar result). In Mann & Repp's data, the size of the shift toward "t" responses after [?] was significantly smaller when the fricative was preceded by a vowel with which it could be syllabified. They hypothesized that listeners would compensate less for coarticulation with a preceding segment that belonged to a different syllable from the stop. We tested the generality of this finding in a third experiment by removing the syllable that had originally preceded the fricative, thereby forcing the preceding fricatives to syllabify with the members of the following stop continuum. We found that English listeners did not respond "t" any more often following an [?] that was not preceded by a vowel than after one that was. This finding, like that obtained in the second experiment from Japanese listeners, indicates that context effects are largely indifferent to whether the target sound and its context are prosodically affiliated, contrary to the findings of both Mann & Repp (1981) and Samuel & Pitt (2003).

We will argue in the general discussion that the first experiment shows that auditory contrast with a preceding fricative influences both Japanese and English listeners' stop place judgments. The results of the second and third experiments support separating a linguistically naïve stage from a later linguistically informed one because such a stage of processing would know nothing about the prosodic affiliation of neighboring sounds.

1 Experiment 1

1.1 Method

1.1.1 Structure of the stimuli: The stimuli in Experiment 1 consist of a sequence composed of the syllable [ʔu] followed by the syllable [suʔ] or [ʔıʔ], which in turn is followed by a syllable beginning with a member of a [k-t] continuum and ending in [o]. The resulting combinations are of the form [ʔusuʔCo] or [ʔuʔıʔCo], where C represents a stop drawn from a t-k continuum. Because the voiceless vowels do not intervene phonetically between the fricative and the stop, we represent these strings henceforth as [ʔusCo] or [ʔuʔCo]. For a variety of reasons, [o] is the best and the only choice for the vowel to follow [t] and [k]. In Japanese, the two high vowels [i] and [u] devoice word-finally after [t] and [k], and furthermore would affricate [t] to [t?] or [ts], respectively. The vowels [e] and [a] would activate actual words because *shite* and *shita* are the respective gerundive and past forms of the light verb *suru* ‘do’, and therefore any *x-shite* or *x-shita* sequences would sound like pseudo-verbs (cf. *kisu-shita* ‘kissed’). Finally, the transitional probabilities between [t] and [k] and a following [o] are more similar than with any other following vowel: in Amano & Kondo’s (2000) lexical database of Japanese, the type frequency ratio of [to] to [ko] is 0.93, or nearly 1.

To test the spectral effects of *un*devoiced vowel contexts concurrently, we include stimuli in which one of the syllables [ni], [nu], or [na] replace the [su?] or [?i?] syllable, yielding stimuli of the form [?uniCo], [?unaCo], and [?unuCo] in addition to [?usCo] and [?u?Co]. The syllable [ni] is a spectrally high and front context, [nu] a spectrally low and back one, and [na] is intermediate in both its spectrum and articulation. Both the auditory contrast and compensation for coarticulation accounts predict that English and Japanese listeners should respond “t” most often after voiced [u], less often after voiced [a], and least often after voiced [i].

The initial syllable [?u] in both the voiceless and voiced vowel stimuli staves off lexical effects, as *suto* ‘strike’ and *niko* ‘two pieces’ are real Japanese words. The materials were also synthesized with an unaccented F0 contour to ensure that they would be perceived as nonce words—the real words *suto* and *niko* are accented on the initial syllable. The unaccented F0 contour consists of a low value on the first syllable, followed by a rise onto the second syllable, which is sustained with slight declination through the end of the stimulus. These efforts were successful in that no Japanese listeners recognized any of our stimuli as real words.

1.1.2 Recording: To obtain the bulk of the phonetic content needed for constructing our stimuli, a native speaker of Tokyo Japanese (the second author) was recorded pronouncing words of the following form: runV_ihV_i where V_i={a, i, u}, and ruCV_jhV_j where CV_j={su, ?, ?i}. 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 the following consonant or the vowel in the next syllable. The frame sentence used was *d?aa* ____ *de onegai* 'please do with X'. The utterances were recorded in a sound-attenuated recording booth to an external desktop computer at a sampling rate of 44100 Hz with 16-bit resolution and encoded as .wav files for later manipulation.

1.1.3 Synthesis and stimulus construction: We extracted the F0, formant, and intensity trajectories for the sonorant portions of selected tokens using Praat (Boersma & Weenink, 2007), and used these values as parameters for resynthesizing the intervals via the Sensyn implementation of KLSYN88 (Klatt & Klatt, 1990). To create a uniform base for the initial [?u] of all the nasal-vowel stimuli, we replaced their values for that interval with the *ru-* values from a token of *runiko*, and resynthesized them. From the midpoint of the nasal onward, the *runiko* token had a rising F2 and F3 into the vowel, whereas *runuko* had a falling F2 and F3, mimicking formant transitions of naturally produced [i] and [u]. *Runako* had an intermediate F2 and F3 contour like that which occurs into the low vowel [a].

Naturally produced [s] and [?] were used for fricatives: the [s] we chose had the highest frequency energy concentration of 10 recorded tokens, while the [?] had the lowest. The spectra of the fricatives are displayed in Figure 1. The spectrum of [?] in Figure 1a has a strong peak just below 4000 Hz and less energy at higher frequencies, while that of [s] in Figure 1b shows increasingly more energy from 4000 to 8000 Hz. The third formant of devoiced [i?] can be seen in the spectral peak at about 3000 Hz in Figure 1a, while the second formant of devoiced [u?] is visible in the spectral peak at about 1800 Hz in Figure 1b.

*** Please put Figure 1 about here. ***

Naturally produced [t] and [k]-bursts taken from utterances of the strings [?uhoto] and [?uhoko] were used as endpoints for a 12-step stop place continuum. Since the [k]-bursts were all longer than the [t]-bursts, we chose the longest [t]-burst, and trimmed a few ms from the end of the [k]-burst to equalize their durations to 15 ms (voiceless stops are unaspirated in Japanese). We then mixed their waveforms together in inversely varying intensities to form the continuum. Across the whole continuum, we boosted the relative intensity of the [k] burst so that in isolation roughly half of the continuum was heard as “t” and half as “k”. The place of articulation of the stop was also conveyed by manipulating the onset frequencies of the [o]’s formants. This vowel lasted 150 ms. In all steps along the [k-t] continuum, F1 was 275 Hz at the onset of the vowel and rose linearly to 475 Hz 70 ms later and remained at that frequency for the remaining 80 ms. At the [k] endpoint, the onset frequencies of F2, F3, F4, and F5 were 900, 2400, 3300, and 4100 Hz, respectively, while at the [t] endpoint, these values were 1450, 2800, 3800, and 4600 Hz, respectively, that is, all higher than at the [k] endpoint. For F3, F4, and F5, these values were sustained for 25 ms before following a 20 ms linear transition to the [o]’s target frequencies of 2600, 3480, and 4193 Hz, respectively. During the remaining 105 ms of the vowel, these formants changed frequency linearly to final values of 2500, 3400, and 4100 Hz, respectively. The F2 transitions were longer and converged later: after 25 ms at their initial values, they fell linearly over a 20 ms interval from 900 to 860 in the [k] endpoint and from 1450 to 973 Hz in the [t] endpoint. Over the remaining 105 ms of the vowel, F2 fell from these values (860 and 973 Hz) to 850 Hz.

Intermediate steps along the continuum were determined by linear interpolation between these endpoint values. The resulting 12 steps were then combined with the 12-step stop burst continuum to create the [ko-to] continuum. These manipulations ensured that the continuum completely spanned the two categories and that the range of ambiguous stimuli was far enough from the endpoints that we could detect any shifts in the category boundaries caused by the contexts. The silent interval lasted 40 ms following voiced vowels, and 60 ms following voiceless syllables.[4] Figure **2**2 displays illustrative spectrograms of our stimuli.

*** Please put Figure 2 about here. ***

In order to show the formant frequencies of the vowels in the stimuli where they are voiced, the highest frequency displayed in the first three spectrograms is 5000 Hz. The frequency ceiling is raised to 8000 Hz to show the differences between the fricative

---

[4] This difference in the duration of the silent interval was unintentional. As reported in the results from Experiment 1, Japanese listeners' stop place judgments were biased by preceding fricatives but not by preceding voiced vowels, despite voiced vowels being closer to the stop than the fricatives. See also the results of Experiment 2, where the duration of the silent interval was manipulated to change the percept of the stop as singleton versus geminate.

spectra more fully in the last two spectrograms. The spectrogram of [?u?i?ko]

1.1.4 Participants: 21 native speakers of Japanese participated in the experiment. They were recruited from International Christian University, the Tokyo University of Agriculture and Technology, and several non-academic venues. All participants received 1,000 yen per hour for their time. One listener gave [ko]-responses for nearly all the stimuli, so her data were excluded from analysis. Results reported here are therefore based on responses from the remaining 20 Japanese participants. 20 adult native speakers of English also participated. They were recruited from the University of Massachusetts, Amherst community. They earned either linguistics course credit or $10/hour for their participation. All participants from both language backgrounds were adults, and they reported no hearing or speaking disorder, and likewise no significant exposure to any language other than Japanese or English before beginning school.

1.1.5 Listening conditions: For the Japanese participants, the experiment took place in a quiet but not sound-attenuated or otherwise special room, while for the English participants, the experiment took place in a sound-attenuated room in the Phonetics Laboratory at the University of Massachusetts, Amherst. The Japanese listeners sat in front of a laptop PC, while the English listeners sat in front of a desktop PC input-output terminal connected to an external computer. All stimuli were output at 16 kHz from the PCs and presented binaurally at a comfortable volume, via Beyerdynamic DT 250 80 Ohm headphones for the Japanese listeners and via sound-isolating Sennheiser HD 280 64 Ohm headphones for the English listeners. Cedrus SuperLab Pro software (version 2.0.4) was used to present all sound stimuli and visual cues, and logged all responses. All the participants used Cedrus RB-834 response boxes to enter their responses.

1.1.6 Trial and block structure: Each test trial began with a single stimulus, representing a point along one of the five continua, played to the listener. When the sound finished playing, two color-coded visual prompts appeared onscreen, to cue listeners to respond. For Japanese listeners, the two prompts took the form of the *katakana* orthography for the syllables “to” and “ko”, and for English listeners, they took the form of “t” and “k”. The prompts were displayed for up to 1500 ms, which was the entire interval during which listeners could enter their response (no responses were collected while the stimuli were being played). After the listener responded or after the 1500 ms interval had elapsed, the software waited an additional 750 ms before presenting the next stimulus.

Listeners started with a short training section that presented 4 repetitions of each continuum endpoint, in random order. The training trials differed from the test trials in a single regard: after listeners responded to each training stimulus, feedback in the form of the correct answer—either “t” or “k”, or the “to” or “ko” kana—would appear onscreen for 500 ms. Following that, the standard 750 ms ITI would pass, and the next stimulus would be presented.

Once listeners finished the training section, they proceeded through 7 test blocks, in which they categorized stimuli drawn from the entire continua. Each block comprised the entire stimulus array, with extra repetitions of the middle steps according to the ratio 3:2:1 for steps {4, 5, 6, 7, 8, 9}, {2, 3, 10, 11}, and {1, 12}, respectively. Therefore, during the entire experiment, participants listened to the endpoints 7 times each, the near-endpoints 14 times, and the middle steps 21 times. Like the training block, all stimuli in the test blocks were presented in random order. Listeners pressed one of two buttons to indicate which consonant they had heard (see instructions below).

After completing each test block, participants received a message telling them to take a short break, and to press a button to continue at their leisure. After completing the fourth block, which marked a near-halfway point, a different message instructed the participants to exit the testing room for a 5-minute break. All experiment sessions were finished within 60 minutes, including time for the reading and signing of consent forms, written and verbal instructions, testing time, breaks, debriefing, and distribution of cash or credit receipts.

1.1.7 Instructions: Before starting the experiment, listeners were told they would hear a variety of syllable-sequences, and to pay attention to the final syllable. Japanese listeners were told to decide whether it sounded like “to” or “ko”, and English listeners whether it sounded like it began with a “t” or “k”. They were told they would start with an initial training section during which they would receive feedback in the form of the correct answer after responding to each trial. They were told all the details of the trial structure: the duration of their response interval (1.5 seconds); that after each stimulus finished playing, two color-coded visual choices would appear—matching the color and arrangement of the two response buttons they were to use on their button boxes; etc. Participants were also instructed to respond as quickly as possible, and to rest their forefingers or thumbs on the two response buttons to allow them to respond quickly without moving their arms or hands.

The only experimental condition concerned the position of buttons for the “t” and “k” response. Half the participants used their left button for the “ t” and right button for “k”, with corresponding visual prompts, and vice versa for the remaining participants. The button conditions compensate for any bias or predisposition participants may have toward either the left or right response. The color coding for button position remained constant across both conditions: the left response was always red, and the right response always blue.

1.2 Results

1.2.1 Japanese listeners: Figure 3a plots the mean proportion of “to” responses for each step along the [t-k] continuum in the three vocalic contexts for the Japanese listeners. Figure 3b plots the total proportion of “to” responses across the entire [t-k] continuum. In Figure 3b, and those that follow, error bars represent 95% confidence intervals. These proportions do not differ between the three vowels ($F < 1$; all statistics are carried out on the total proportions of “t” responses across the entire continuum). Button also did not affect responses ($F < 1$), nor did it interact with vowel context ($F(2, 36) = 1.32$, $p = .28$).

*** Please put Figure 3 about here. ***

Figure 4a shows the mean proportions of “to” responses obtained from the Japanese listeners in the two fricative contexts. Figure 4b shows the mean total proportions of “to” responses across the entire continuum for these listeners.

[?] induced significantly more "t" responses than [s] (*F*(1, 18) = 9.05, $p < .01$). Neither button nor its interaction with the fricative place difference were significant ($F < 1$ and *F*(1, 18) = 1.77, $p = .2$, respectively).

*** Please put Figure 4 about here. ***

1.2.2 English listeners: Figure 5a shows the mean proportions of "t" responses obtained from the English listeners to each step along the [k-t] continuum in the three vowel contexts. Figure 5b shows the mean total proportions of their "t" responses across the entire continuum.

*** Please put Figure 5 about here. ***

English listeners responded "t" most often after [u], less often after [a], and least often after [i]: the proportion of "t" responses correlates negatively with the preceding vowel's F2 and likewise with its backness. 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 [a] and [i] did not even reach marginal significance ($F < 1$). Neither button nor its interaction with vowel was significant

(*F*(1,18) = 1.3, *p* = .269; *F*(2,36) = 1.824, *p* = .176).

Figures 6a and 6b show the effects of preceding fricatives on the identification of the [t]-[k] continuum for English listeners. Listeners responded “t” significantly more often after [?] than [s] (*F*(1,18) = 6.798, *p* = .018). Listeners also responded “t” significantly more often when they used their left button for the “t” response rather than their right button (0.515 vs 0.391; *F*(1, 18) = 4.455, *p* = .045), but button did not interact significantly with the fricative quality (*F* < 1).

*** Please put Figure 6 about here. ***

1.3 Summary and discussion

The preceding fricative but not the preceding vowel influenced Japanese listeners’ stop place judgments, while both the fricative and the vowel influenced the English listeners’. That a preceding [?] caused Japanese as well as English listeners to respond “t” more often than a preceding [s] confirms the prediction of the auditory contrast account (Hypothesis 1) that listeners would hear a stop that is spectrally intermediate between [t]’s high value and [k]’s low one as contrasting with the spectrum of its fricative context. In other words, the spectrally high stop response, “t”, is given more often after the spectrally low fricative [?] than after the spectrally high [s].

It also disconfirms the predictions of two of the alternative versions of the compensation for coarticulation, Hypotheses 2a and 2b. Hypothesis 2a predicts counterfactually that listeners should respond “t” more often after [s] than [?] because the devoiced vowel is [u?] when the fricative is [s]. Hypothesis 2b predicts that they should instead respond “t” no more often after [s] than [?] because they compensate for coarticulation with [s] as well as [u?], and the two perceptual adjustments cancel one another out. The only version of the compensation for coarticulation account that is not disconfirmed by these results is Hypothesis 2c, in which listeners compensate for coarticulation with the fricative alone. However, as noted in the introduction, accepting this hypothesis as an account of the results presented here requires us to amend the compensation for coarticulation account to permit listeners to attend to just one coarticulatory effect and entirely ignore the other. We present evidence in the general discussion that stops do coarticulate with preceding voiceless vowels as well as and to nearly the same extent as preceding fricatives. Because listeners have acoustic evidence that should lead them to compensate for coarticulation with the voiceless vowels, it is surprising that they do not do so.

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. A moment’s thought leads to the conclusion that 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, why do Japanese listeners not also compensate similarly? Third, as we will show later, 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 failure of the Japanese listeners to compensate for the voiced vowels is also a challenge to 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. Nonetheless, the auditory contrast account is only challenged by this failure of the voiced vowels to influence Japanese listeners’ stop place judgments, while the compensation for coarticulation account is also challenged by Japanese and English listeners’ failure to compensate for coarticulation with the voiceless vowels.

Finally, like Hypothesis 2a, the phonological contrast account predicts counterfactually that Japanese listeners would respond “t” more often after [s] because these listeners would infer from their phonological knowledge that the vowel [u̥] intervenes between [s] and the stop. In disconfirming the phonological contrast account, these results also show that sequential contrast operates before Japanese speakers hallucinate a high vowel between the fricative and stop (cf. Dupoux, et al., 1999, 2001). This supports the hypothesis of an initial stage of processing which is impervious to the influence of linguistic knowledge.

Whatever the finding that a preceding vowel does not influence Japanese listeners' stop place biases may mean for testing the competing hypotheses, we must still try to explain its lack of influence. We try to do so in the second experiment.

2 Experiment 2

One potential reason why a preceding vowel may have influenced English but not Japanese listeners' stop place judgments is that the English listeners would have syllabified the stop with the preceding vowel but the Japanese listeners would not.

On the one hand, English phonology permits a single intervocalic stop to be ambisyllabic—belonging both to the coda of the preceding syllable and the onset of the following one (Kahn, 1976), as shown in Figure 7b. English listeners would have been encouraged to parse the stop into both syllables by the rise in F0 on the second syllable, which would have led them to interpret that syllable as stressed. Because stressed syllables attract consonants (Kahn, 1976; Kirk, 2001), interpreting the second syllable as stressed would in turn have led English listeners to parse the consonant into that syllable as well as the following one.

On the other hand, Japanese phonology requires a single intervocalic stop to be exclusively an onset, regardless of the preceding context (Figure 7a). When the preceding syllable's nucleus is a voiced vowel, that syllable is open. When the preceding syllable's nucleus is instead a devoiced vowel, the resulting syllable is "defective", in that it is not parsed into the next larger prosodic constituent, the accentual phrase. The defective status of such syllables is independently motivated by their inability to carry pitch accents (Haraguchi, 1977; McCawley, 1977; Sugito & Hirose, 1988). In syllables with devoiced nuclei beginning with fricatives, like the ones considered here, the fricative also cannot be syllabified with the *preceding* syllable, because Japanese never allows coda consonants that are not homorganic with the following consonant (Itô, 1986, 1989), as shown in Figure 7c). The syllabification of these strings by English listeners depends on whether they notice the presence of the devoiced vowel. If they do, it could serve as the nucleus of a syllable whose onset would be the voiceless fricative. If they do not, the fricative could syllabify as the coda of the preceding syllable (Figure 7d) or join the stop in the following syllable's onset (not shown). These considerations predict that the sequences [?uniko] and [rusuko] would be parsed into syllables by Japanese and English listeners as in Figure 7.

*** Please put Figure 7 about here. ***

Given the syllabification patterns in Figure 7, to explain the patterns in Experiment 1, we can entertain the hypothesis that auditory contrast does not operate across a syllable boundary. Finding an effect of syllabification would indicate that knowledge of the prosodic grouping of segments influences perceptual processing from the outset, contrary to the prediction of the autonomy hypothesis. This result would support the contention of Dupoux, et al. (2000, 2001) that language-specific phonotactics immediately separate the constituents of incoming signals into syllables. We tested this hypothesis in Experiment 2 by lengthening the duration of the silent interval corresponding to the stop closure from 80 to 250 ms. Doing so would cause Japanese listeners to perceive the stop as geminate rather than a singleton. This percept would force them to syllabify the resulting geminate stop into the preceding coda as well as the following onset, because the first part of a geminate consonant belongs to the preceding syllable in Japanese and the second part to the following syllable (Itô, 1986, 1989).

2.1 Method

2.1.1 Stimuli: The stimuli consisted of the syllable [ʔu] followed by one of {ni, na, nu}, which was in turn followed by the [ko-to] continuum. One stimulus subset was identical to that used in Experiment 1, while in the other, the silent interval between the end of the second vowel and the release of the stop was increased from 80 to 250 ms to create the percept of a geminate rather a singleton stop.

2.1.2 Procedures: With three exceptions, procedures were identical to those used in Experiment 1. First, 21 native speakers of Japanese participated. Second, the listeners sat in the sound-attenuated room in the Phonetics Laboratory of the University of Massachusetts, Amherst and listened to the stimuli through sound-isolating Sennheiser HD 280 64 Ohm headphones. Finally, listeners chose from four rather than just two responses: “t”, “tt”, “k”, and “kk”. The only responses analyzed were those where the singleton vs geminate choice corresponded to the duration of the silent interval. Listeners responded “t” or “k” to 97.6% of trials in which the silent interval was short, and “tt” or “kk” to 96.1% of trials when it was long, so nearly all their responses were included in the analyses. Each endpoint stimulus (1, 12) was presented a total of 18 times to each listener, the near endpoint stimuli (2, 11) 36 times, and the intermediate stimuli (3-10) 54 times.

2.2 Results

Figures 8-9a,b are plots of the mean proportions of “t” and “tt” responses to each stimulus from the [k-t] and [kk-tt] continua, respectively, as a function of the preceding vowel. The figures show that the preceding vowel does influence stop place judgments differently when the silent interval was long than when it was short. However, they show that only the influence of [a] changes with the duration of the silent interval; listeners respond “t” more often after [a] than [i] or [u] (Figures 8a,b), but they respond “tt” less often after this vowel than the other two (Figures 9a,b). Otherwise, listeners respond “t” no more often after [u] than [i] for the long than the short duration. Comparison of the two figures also shows that listeners responded “t” more often to the stimuli with short silent intervals than they responded “tt” to those with long silent intervals.

*** Please put Figures 8 and 9 about here. ***

A repeated measures ANOVA on the total proportion of “t” or “tt” responses shows that “t” responses are significantly more frequent than “tt” responses (.548 vs .490, $F(1,19) = 34.127$, $p < .001$). Vowel did not influence responses by itself ($F(2,38) = 2.806$,

$p = .076$) but it did interact significantly with consonant length ($F(2,38) = 15.864$, $p < .001$). Separate analyses of the responses to the singleton and geminate stimuli show that this interaction reflects significantly more frequent singleton “t” responses following [a] than [i] ($t(20) = 3.723$, $p = .001$) or [u] ($t(20) = 3.612$, $p = .001$) but no difference in geminate “tt” responses following [a] than [i] ($t(20) = 1.196$, $p > .10$) or [u] ($t(20) = 1.042$, $p > .10$). The effect of button assignment was significant ($F(1, 19)=6.258$, $p=.022$). Listeners who pressed right buttons for “t” and “tt” responses gave these responses more often than those who pressed left buttons (“to/tto” -right=.540; “to/tto”-left=.496), but button did not interact with consonant length nor vowel (button by length: $F(1, 19) = 1.583$, p > .10; button by vowel=$F < 1$).

2.3 Summary and discussion

Listeners identified the stop as [t] rather than [k] more often when the stop was a singleton than a geminate, and they also identified the stop as [t] more often after [a] than [i] or [u] when the stop was a singleton. However, lengthening the silent interval enough to get listeners to hear the stop as a geminate and thereby syllabify it with the preceding vowel had no effect on how likely a listener was to identify the stop as [tt] following [u] compared to [i]. This last finding shows that the failure to obtain any effect of vowel in Experiment 1 cannot be explained by the fact that Japanese listeners would parse the vowel and a singleton stop into separate syllables. Also, the single effect of vowel in this experiment appeared in responses to the singletons, not to geminates. These results therefore indicate that syllabification is not responsible for the lack of an effect of vowel on stop place judgments in Japanese listeners' responses. These results also indicate that the affiliation of the stop to the preceding vowel context does not alter that context's (lack of) effect, contrary to what we would expect if the signal's constituents are immediately separated into syllables by the language's phonotactics (cf. Dupoux, et al., 1999, 2001). In the next section, we present the results of another experiment designed to test whether syllable affiliation influences the context's effects on stop place judgments. Finally, the bias toward more 't' responses than 'tt' responses may be due to the Japanese suffix -*kko* 'people', which could have biased the listeners against "tt" and toward "kk" responses when the stimuli were geminates.

3 Experiment 3

In Experiment 2, we failed to find evidence that the strength of the perceptual biases caused by a neighboring sound depends on that sound belonging to the same syllable as the one whose perception it biases. Mann & Repp (1981), however, found that the effect of a preceding [s] vs [?] on judgments of a following [k-t] continuum was significantly larger when no vowel preceded the fricative than when one did. They suggested that when the vowel is present, listeners would syllabify the fricative with it, and in doing so would be less likely to treat the following stop as coarticulated with the fricative, or at least as less strongly coarticulated with that fricative. If the coarticulation were perceived to be less likely or weaker when the fricative belongs to another syllable, then listeners would compensate less and their judgments of the stop’s place would differ less as a function of the fricative’s place. Samuel & Pitt (2003) also present evidence suggesting that the strength of the effect of the fricative’s place on the judgment of the following stop’s place varies inversely with how strongly the fricative is affiliated with the preceding string.

In Experiment 3, we tested the generality of Mann & Repp’s and Samuel & Pitt’s findings by removing the initial syllable [?u] from the stimuli with devoiced vowels and presenting those stimuli to English listeners, along with the original stimuli, for identification of the stop’s place. If Mann & Repp’s results generalize, then the difference in “t” responses between the [s] and [?] contexts should be greater in the stimuli from which the syllable [?u] has been removed than in the original stimuli.

3.1 Method

The stimuli in this experiment were constructed by simply removing the syllable [?u] from the original devoiced vowel stimuli by means of digital waveform editing. These stimuli were presented together with the original devoiced vowel stimuli for identification of the stops' place. Otherwise, methods and procedures were identical to those in Experiments 1 and 2. Endpoint stimuli (1, 12) were presented 14 times to each listener, near endpoint stimuli (2, 3, 10, 11) 28 times, and intermediate stimuli (4-9) 42 times. Responses were collected from 22 English listeners.

3.2 Results

Figures 10-11a,b show the mean proportions of "t" responses for each stimulus as a function of the preceding fricative for the original stimuli with the preceding syllable [?u] (Figures 10a,b) and for those without [?u] (Figures 11a,b). The figures show that listeners responded "t" more often after both [?u?] and [?] than after [?us] and [s]; they also hint at a greater difference after the fricatives not preceded by [?u]. A repeated measures ANOVA shows, however, that only the effect of fricative was significant ($F(1, 19) = 21.594$, $p < .001$), and that neither the presence vs absence of [?u] nor its interaction with the fricative significantly influenced the frequency of "t" responses ($F(1, 19) = 1.151$, $p > .10$; $F < 1$, respectively). Button did not significantly affect responses ($F(1, 19) = 2.759$, $p > .10$) nor did it interact with fricative or the presence vs absence [?u] (both *F*s < 1).

*** Please put Figure 11 about here. ***

3.3 Summary and discussion

Unlike either Mann & Repp (1981) or Samuel & Pitt (2003), we did not find that the presence of a preceding string with which the fricative might prosodically affiliate significantly diminished the size of the fricative’s effect on English listeners’ judgments of a following stop’s place: the difference in “t” proportions between the two fricative contexts is .097 when [?u] is absent and .088 when it is present. These results indicate that the affiliation of the fricative context to preceding sounds does not alter its influence on the stop place judgment.

4 Summary and general discussion

4.1 Summary

Experiment 1 showed that both Japanese and English listeners’ judgments of the place of articulation of a stop drawn from a [k-t] continuum depend on the preceding fricative: both groups of listeners respond “t” more often after [?] than [s]. Beside the similar fricative effect on the responses of the two groups of listeners stands a sharp difference in the effects of a preceding vowel: Japanese listeners’ stop place judgments are unaffected by this vowel, while English listeners respond “t” more often after [u] than [a] and more often after [a] than [i].

Experiment 2 tested the hypothesis that the preceding vowel did not affect Japanese listeners’ stop place judgments because the phonotactics of Japanese force the vowel and the stop into separate syllables but into the same syllable in English. It did so by lengthening the silent interval corresponding to the stop closure enough to make Japanese listeners hear the stop as a geminate rather than a singleton. Because geminates are syllabified with the preceding as well as the following vowel, a long closure should cause a preceding voiced vowel to influence stop place judgments if that influence depends on the vowel belonging to the same prosodic constituent as the stop. This manipulation did alter [a]’s effect on stop place judgments—listeners responded “t” more often after [a] than [i] or [u] when the silent interval was short enough to convey a singleton stop closure but not when it was long enough to convey a geminate instead. Otherwise, “t” responses were not any more frequent after [u] than [i] when the silent interval was long enough to convey a geminate stop closure than when it was short enough to convey a singleton stop closure instead.

Experiment 3 tested the perceptual effect of affiliation in another way, by removing the initial syllable [?u] to see whether the fricative’s influence on English listeners’ stop place judgments was greater when no vowel preceded with which it could affiliate, as predicted by Mann & Repp's (1981) and Samuel & Pitt's (2003) results. The results indicated quite clearly that it was not.

In the next two sections, we discuss the results obtained in Experiment 1 in terms of the competition between the various hypotheses outlined in the Introduction. The first of these sections discusses the auditory vs phonological contrast hypotheses (Hypotheses 1 vs 3), while the second discusses the auditory contrast vs compensation for coarticulation hypotheses (Hypotheses 1 vs 2a,b,c). These discussions are followed by a report of an analysis of the acoustic effects of coarticulation between stops and preceding fricatives and voiceless vowels. The purpose of this analysis is to establish that stops coarticulate with both preceding segments, and thus make a *prima facie* case for compensation for articulation with both, as predicted by Hypothesis 2b. We close with discussion of the failure of preceding voiced vowels to influence stop place judgments, and the failure of the two experiments that tested the hypothesis that the context's perceptual effect depends on its being prosodically affiliated with the target sound.

4.2 Auditory vs phonological contrast

In Experiment 1, the percept of the stop could contrast with the percept of the sound that occurs next to it, physically or phonologically. In Japanese [s{k-t}] and [?{k-t}] sequences, the fricatives [s] and [?] physically occur next to the following stop,[5] while the vowels /u/ and /i/ occur next to it phonologically. If it is the physically adjacent fricative that determines responses, then Japanese listeners are predicted to respond “t” more often after [?] than [s]. If the Japanese listeners instead infer that the devoiced vowel is present immediately before the stop because they know that the strident fricative is pronounced [?] before devoiced [i?] but [s] before devoiced [u?], that is, if it is the phonologically adjacent sound that serves as the context for perceiving the stop, then they are predicted to respond “t” more often after [s].

Finding an effect of the physically adjacent fricative would support the more general hypothesis advanced by Kingston (2005) that sequential contrast arises during an autonomous, linguistically naïve, auditory processing stage, while finding an effect of an inferred phonologically adjacent sound would instead indicate that sequential contrast arises between phonological rather than phonetic categories and thereby provides no grounds for separating auditory processing from a stage when linguistic knowledge shapes listeners’ percepts.

The results of Experiment 1 indicated that the relevant context was the adjacent

---

[5] For the sake of limiting the comparison between hypotheses in this section to the auditory and phonological contrast hypotheses, we will assume during this section that the devoiced vowel’s oral articulation is either not present in the fricative or that it is not perceptible if it is there. In the next section, we reexamine this assumption.

fricative rather than an inferred vowel and therefore supported the hypothesis that the contextual effect is auditory rather than phonological. In doing so, it also revealed the autonomy of the initial processing of the signal from the influence of linguistic knowledge.

4.3 Auditory contrast vs compensation for coarticulation

The effects of [s] vs [?] on judgments of a following [k-t] continuum had originally been explained by Mann & Repp (1981) as compensation for coarticulation: listeners would identify a stop intermediate between [k] and [t] as “t” more often after [?] than [s] because they undid the expected backing of the stop by the palatoalveolar fricative. That account would have predicted the effect of the preceding fricative on English listeners’ stop place judgments in Experiment 1, but only if they did not perceive a vowel during the fricatives. As originally laid out in the statement of Hypothesis 2a, English listeners are predicted to perceive and compensate for the coarticulation with the devoiced vowel just as much as Japanese listeners are, so the English listeners’ apparent failure to do so would also require special explanation. There is even less reason to expect the compensation account to rule out an effect of the voiceless vowel on Japanese listeners’ stop place biases as their linguistic experience should aid them in detecting its presence in the signal and its influence on the following stop’s acoustics. In short, there is every reason to expect Japanese listeners to perceive a devoiced [u?] in the [s] and a devoiced [i?] in the [?].

They would do so because Japanese speakers still produce the oral articulation of the devoiced vowels, and Japanese listeners can perceive its quality during the fricative (Beckman & Shoji, 1984; Tsuchida, 1994; Nakamura, 2003). If the Japanese listeners were compensating for coarticulation with the voiceless vowel alone, those vowel percepts should have led them to hear more “t”s after [s] than [?], because the vowel following [s] is back [u?]. This predicted result is precisely the opposite of the one obtained. This version of the compensation for coarticulation account (2a) makes the same predictions about how the fricative contexts should alter the Japanese listeners’ responses as the phonological contrast account and is thus disconfirmed by the same evidence.

The second version of the compensation account, Hypothesis 2b, predicts that listeners would also compensate for coarticulation with the fricative and thereby hear more “t”s after [?] than [s]. This compensation would have exactly the opposite effect of compensating for the voiceless vowels that condition the difference in fricative place, and the two perceptual adjustments should cancel one another’s effects on stop place biases. The predicted result, no contextual bias, was also not obtained, which disconfirms Hypothesis 2b.

Both versions of the compensation for coarticulation account assume that listeners are compensating for the voiceless vowels' coarticulatory effects, and both were disconfirmed by their apparent failure to do so. This failure raises the following questions: (1) is the devoiced vowel's articulation present and perceptible during the fricative's pronunciation, (2) do 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? For either Hypotheis 2a or 2b to be tenable, the answers to the first two questions must be "yes", and for Hypothesis 2b to be so, the answer to the third must also be "yes". A positive answer to the first question ensures that listeners have independent evidence of the devoiced vowel's articulation to which any coarticulatory effects in nearby segments can be attributed. Such an answer to the second ensures that there is perturbation of a nearby segment's articulation to compensate for. Finally, answering "yes" to the last question ensures that compensation for coarticulation with the devoiced vowel can at least compete with compensation for coarticulation with the fricative.

Negative answers to the any of these questions raises the alternative possibility that the stops only coarticulate perceptibly with the preceding fricatives, that is, Hypothesis 2c. If that is the case, then the compensation for coarticulation account predicts the same stop place biases as the auditory contrast account. If these questions are answered positively, then the compensation for coarticulation account must be amended to explain why listeners compensate for some but not all perceptible coarticulatory effects.

As originally stated, the compensation for coarticulation hypothesis predicted that English listeners should compensate for coarticulation with the devoiced vowels just as much as Japanese listeners do. This outcome is expected if at least the first two questions are answered positively. Positive answers are perhaps more likely for Japanese than English listeners, who may be less sensitive to the acoustic evidence in the fricative and stop of the devoiced vowel's articulation because they lack the Japanese listeners' phonological knowledge of the relationship between the fricative's place of articulation and the quality of the devoiced vowel. Indeed, they lack the knowledge that a devoiced vowel must be present between the fricative and the stop. Moreover, unlike in Japanese, alveolar and palatoalveolar fricatives are not predictable allophones but instead contrasting phonemes in English, which should heighten English listeners' sensitivity to the difference between the fricatives. Taken together, these facts raise the possibility that English listeners would not notice the devoiced vowel, but Japanese listeners would.

Results reported in the literature indicate that the oral gesture of a devoiced vowel is still produced and that adjacent consonants coarticulate with that gesture. In this section, we evaluate that evidence in more detail; in the next we present new evidence of our own. The evidence in the literature is in fact mixed. On the one hand, Nakamura's

(2003) electropalatographic (EPG) evidence shows that the fricative and stop articulations in a sequence where the intervening vowel devoices, for example, in [su?k], are closer together than in one where the vowel remains voiced. The occlusion of [k] can begin before the articulatory offset of [s] in [su?k], although the extent of gestural overlap varies considerably. Nakamura nonetheless finds traces of the oral articulation of the devoiced vowels in his EPG data, which is not surprising given Beckman & Shoji's (1984) and Tsuchida's (1994) evidence that the vowels' oral articulations can be perceived. The following stop could therefore coarticulate with these traces of the vowel articulation as well as with the preceding fricative, and Japanese listeners could compensate for that coarticulation. On the other hand, Tsuchida (1994) found that a fricative also coarticulates with the next consonant when the intervening vowel devoices: the lower edge frequency of the noise in [?] is significantly lower when the following stop is [k] than when it is [t]. This finding indicates that the stop and fricative do coarticulate with one another, but it does not rule out simultaneous coarticulation with the devoiced vowel. However, Tsuchida's evidence does not show whether the stop coarticulates with the devoiced vowel nor if it does whether that coarticulation changes the stop's acoustics in the opposite direction from coarticulation with the fricative, as we might expect from the fact that the fricative's articulation is front before a back devoiced vowel and back before a front one. If, as we suppose, the acoustic effects of coarticulating with the two segments are opposite in direction but equal in size, then we would expect them to cancel out perceptually and listeners not to be biased by context.

To address these uncertainties, we carried out an acoustic study of coarticulation between the segments in these sequences. The results are reported in the next section.

4.4 Coarticulation of stops with preceding fricatives and voiced and voiceless vowels

Four adult speakers of Tokyo Japanese, two men and two women (one of the male speakers is the second author of this paper), were recorded. They produced utterances identical in all but one case to the materials in the perceptual experiments reported above. None of these speakers reported any hearing or speaking disorder. All were students at the University of Massachusetts, Amherst at the time of the recording, but they all reported that they still used Japanese daily for extended periods of time. The recording session lasted about 20 minutes for each speaker, and all speakers but the second author were recompensed $5 for their time.

The portion of interest in these utterances consisted of three syllables: (1) [?u], (2) {na, nu, ni, su??i??u?, and (3) {to, ko}, which was pronounced in the frame *d?aa* ___ *de onegai* “Please do with ___”. The one sequence that is not identical to any of the perceptual stimuli is that in which the second syllable is [?u?. This is the pronunciation of the underlying sequence /sju/ when followed by a voiceless obstruent; that is, /s/ is pronounced as a palato-alveolar fricative before the palatal glide [j] as well as before the high front vowel [i]. This additional second syllable was included in order to test the effect of the devoiced vowel's quality separately from that of the fricative's place. These three-syllable sequences were pronounced as unaccented words so as to match the stimuli used in the perceptual experiments.

These sequences were written on the cue cards in katakana orthography, which is used for nonce and loan word spelling. The frame sentence was not written on the cue cards. The cue cards were shuffled before each reading to randomize the order in which they were spoken. Ten repetitions of each utterance were recorded, but the first was discarded as warm-up, and only the final nine were analyzed. The speakers were encouraged to speak in a casual style, as though to a friend, to ensure that the high vowels were devoiced after the voiceless fricatives. This encouragement was successful: all vowels following fricatives were devoiced.

The speakers wore a head-mounted microphone, whose output was immediately amplified and digitized at 44100 kHz with 16 bit resolution and stored as .wav files as the utterances were being pronounced.

Using conventional acoustic and visual criteria from the waveform and spectrogram displays in Praat (Boersma & Weenink, 2007), we located the stop burst and the onset of voicing in the following [o] in each token. In the stop burst's spectrum, we measured the center of gravity and standard deviation within a frequency band extending from 750-6000 Hz, and at the onset of voicing, we measured the frequency of the second formant (F2).

In calculating the center of gravity and standard deviation of the stop burst's spectrum, the interval within a gaussian window 25 ms long centered on the burst was extracted, band-pass filtered between 750-6000 Hz, and the spectrum calculated via an FFT—all calculations were also carried out using Praat. The delay in the onset of voicing was slightly but significantly longer following the release of [k] than [t] (24 ± 1 ms vs 21 ± 1 ms; $F(1, 469) = 19.501$, $p < .001$), but both average delays in voice onset are long enough that little if any of the voiced portion of the vowel would have been included in this window. This interval was band-pass filtered between 750-6000 Hz because this range is roughly that of F2-F6.

The spectrum's center of gravity (COG) estimates how high or low energy is concentrated, while its standard deviation (SD) estimates how broadly energy is distributed across frequencies. Both values are expected to be larger in the acute, diffuse spectrum of the coronal stop [t] than the grave, compact spectrum of the dorsal stop [k] (Jakobson, Fant, & Halle, 1951; Stevens & Blumstein, 1978; Blumstein & Stevens, 1979).

The F2 frequencies were extracted using Praat's LPC algorithm, with the number of formants set to 5 for all speakers and the maximum frequency of the formants set to 5000 Hz for the male speakers and to 5500 Hz for the female speakers.

Both measures of coarticulation, the burst's spectrum and F2's onset frequency, reflect the stop's place of articulation and the following vowel, as well as any coarticulation with the preceding fricative and/or vowel. They differ principally in how open the vocal tract is, and thus in whether any resonances of the cavity behind the constriction radiate with sufficient intensity as to influence the measurements. The oral cavity is still nearly closed when the stop is released, so back cavity resonances are unlikely to be strong enough to influence the spectrum of the sound radiating from the mouth much. The cavity has opened substantially by the time voicing begins, so back cavity resonances are likely to contribute along with front cavity resonances to the spectrum of the radiated sound at that time.

4.4.1 Statistical analyses: The two measures of the burst's spectrum and F2 were each analyzed statistically in three ways, which differed in which contexts were examined and consequently in what the independent variables were. The first analysis contrasted voiced [ni, nu] and voiceless [?i?, su?] contexts, that is, the contexts which contrasted in the perceptual experiments reported above. These contexts are represented by the independent variables voice (voiced/voiceless) and vowel (i/u). The second analysis contrasted voiced [ni, nu] with voiceless [?i?, ?u?] contexts. Unlike the first analysis, this one unconfounds the effect of vowel backness from fricative place. The same independent variables represented the four contexts in this analysis, too. Finally, the third analysis contrasted voiceless [?u?, su?]. This last analysis unconfounds the effect of fricative place from vowel backness—the voiced contexts were left out of this last analysis because there is no comparable difference in place before the voiced vowels. Context is represented by a single independent variable, fricative place (?/s) in this analysis. In addition to these independent variables, two others, speaker[6] and stop place (t/k),[7] were included in all analyses.

4.4.2 Center of gravity of the stop burst's spectrum: In the analysis of the center of gravity measure in the contexts [ni, nu, ?i?, su?], the effect of voice was significant (*F*(1,269) = 7.289, *p* = .007), but the effect of vowel was only marginally significant

---

[6] Although the speakers differed from one another in various ways, these differences are not of particular interest here and will not be discussed. This variable is instead only used to soak up variance in the dependent variable.

[7] Stop place always has a significant effect on the dependent measures, because center of gravity, standard deviation, and F2 onset frequency are all always substantially higher when the stop is [t] rather than [k]. Therefore, stop place will be discussed only when it interacts significantly with the contextual variables.

($F(1,269) = 3.217$, $p = .074$), and both variables interacted significantly with stop place (voice by stop place: $F(1,269) = 8.175$, $p = .005$; vowel by stop: $F(1,269) = 19.902$, $p < .001$). The three-way interaction between voice, vowel, and stop place was, however, not significant (F < 1, Figure 12a). The figure shows that the difference in the center of gravity of the burst spectrum between [k] and [t] is larger after [u] than [i] for both voiced and voiceless vowels. The voice by stop place interaction is significant because the center of gravity values for [t] but not [k] are lower after voiced than voiceless vowels.

The analysis with contexts [ni, nu, ?i?, ??u?] unconfounds vowel backness from fricative place. Once these properties are unconfounded, voice is still significant ($F(1,277) = 4.041$, $p = .045$), but vowel is no longer even marginally significant ($F(1,277) = 1.505$, $p > .10$). Vowel nonetheless interacts significantly with stop place ($F(1,277) = 4.125$, $p = .043$), and the three-way interaction between voice, vowel, and stop place is significant ($F(1,277) = 9.942$, $p < .001$, Figure 12b). The figure shows that when fricative place is held constant, voiceless [i?] lowers both [k]’s and [t]’s centers of gravity compared to voiceless [u?], but voiced [i] shrinks the difference in center of gravity between [k] and [t] compared to voiced [u]. As the voiced vowel contexts are the same as those in Figure 11a, it is entirely unsurprising that they should have the same effects.

Finally, in the analysis of contexts [?u, ?su?], which unconfounds fricative place from vowel backness, fricative place does not significantly affect the center of gravity of the stop burst’s spectrum (F < 1) on its own, but this variable does interact significantly with stop place (F(1,134) = 4.973, p = .027, Figure 12c). This interaction is significant because the difference in center of gravity between [k] and [t] is larger after [s] and smaller after [?].

*** Please put Figure 12 about here. ***

Figure 12a shows that the difference in center of gravity between [k] and [t] is larger after [su?] than [?i?]. Figure 12c shows the same difference, and moreover shows that the apparent effect of voiceless vowels observed in Figure 12a is most likely an influence of the fricative’s place because the vowel’s backness is held constant in the contexts compared in Figure 12c. A comparison of the voiced vowels’ effects in Figure 12b with the fricatives’ effects in Figure 12c shows that voiced [u] and [s] both increase the size of the center of gravity difference between [k] and [t] compared to voiced [i] and [?].

By this measure of coarticulation with the preceding context, the front fricative and the voiced back vowel have the same coarticulatory effects on the stops’ acoustics, as do the back fricative and the front voiced vowel. Contrary to appearances, the effects of these place differences between voiced vowels and fricatives are not contradictory. The tongue constricts the oral cavity at nearly the same location in the back fricative and the front vowel, between the alveolar ridge and the palate in the fricative and at the front of the palate nearby in the vowel. Moreover, both of these constrictions fall between the constriction locations for [k] and [t], so coarticulation with either one should pull the former forward and the latter backward, and in doing so contract the difference between their centers of gravity. While coarticulation with [s] would pull [k] forward and coarticulation with [u] would pull [t] backward, coarticulation with [s] keeps [t] front and coarticulation with [u] keeps [k] back. As a result, center of gravity differences between [k] and [t] do not contract in these contexts compared to after the complementary contexts [?] and [i]. This parallelism between fricative and vowel contexts is striking because [u] is the vowel that causes the strident fricative to be pronounced [s], and [i] is the vowel that causes it instead to be pronounced [?].

When fricative place is instead held constant, as in Figure 12b, and only vowel backness differs, voiceless vowels have a different effect than voiced ones. After the voiceless vowels, the difference in center of gravity is about the same size after voiceless [i?as voiceless [u?, and [i?lowers both stops' centers of gravity compared to [u?. This effect is also not large. In an ANOVA on just the voiceless vowel contexts alone, [?i,??u?, neither the effect of vowel nor its interaction with stop place are significant (both *F*s < 1). This outcome differs from that obtained when the voiced contexts [ni, nu] are examined alone. Vowel is not significant by itself (*F*(1,141) = 1.005, *p* > .10) but does interact significantly with stop place (*F*(1,141) = 12.361, *p* = .001). This pair of results indicates that by the center of gravity measure, stops coarticulate noticeably with preceding voiced vowels but not preceding voiceless ones.

More generally, the analysis of this measure of the stops' coarticulation with preceding fricatives and vowels shows that, contrary to expectation, neither [s] and [u] nor [?] and [i] have opposite effects on the following stop's acoustics. We had expected that the more posterior sounds in each pair, [?] and [u], would pull both the stops' articulations back and by lengthening the cavity in front of the constriction lower the centers of gravity in the burst spectra compared to the more anterior sounds, [s] and [i]. Instead, what we observe is that the anterior fricative and the posterior vowel both prevent the center of gravity difference between [k] and [t] from contracting compared to the posterior fricative and the anterior vowel.

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 after [i] as [?]. 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 only compensate for coarticulation with the preceding fricatives, as there is no perceptible acoustic evidence of coarticulation with the voiceless vowels. The first prediction is disconfirmed by both English and Japanese listeners' responses. The English listeners responded "t" more often after voiced [u] but less often after [s] compared to voiced [i] and [?], and Japanese listeners' stop place biases did not differ after voiced [i] compared to [u]. The second prediction fares somewhat better as both English and Japanese listeners apparently compensated for coarticulation with the fricative alone in the voiceless vowel contexts.

4.4.3 Standard deviation of the stop burst's spectrum: We turn next to the analyses of the burst spectrum's standard deviation, the measure of how diffuse or compact it is. In the analysis where the contexts are [ni, nu, ?i? su?, neither voice nor vowel has any significant effect by itself on the standard deviation (both *F*s < 1), nor are any of their interactions with stop place any more than marginally significant (voice by stop: *F*(1,269) = 3.127, *p* = .078; vowel by stop: *F*(1,269) = 2.367, *p* > .10, voice by vowel by stop *F*(1,269) = 2.127, *p* > .10, Figure 13a). All the figure indicates with any confidence is that the standard deviation is, as expected, substantially larger for diffuse [t] than compact [k].

The analysis of the contexts [ni, nu, ?i??u?, which unconfounded vowel backness from fricative place, yields clearer effects: voice is just marginally significant (*F*(1,277) = 2.779, *p* = .097) and does not interact with stop place (*F* < 1), but vowel does reach significance (*F*(1,277) = 4.458, *p* = .036) and interacts significantly with stop place

($F(1,277) = 15.557$, $p < .001$). The three-way interaction is not significant ($F(1,277) = 1.208$, $p > .10$, Figure 13b). The two-way interaction between vowel and stop is significant because the difference between [k]’s compact burst and [t]’s diffuse one is larger after [i] than [u]. Notice that the voiceless vowels affect the difference between the [k] and [t] bursts’ standard deviations at least as much as if not more than the voiced ones do.

In the analysis of the contexts [?u,?su?], where fricative place is unconfounded from vowel backness, the effect of fricative place was only marginally significant ($F(1,134) = 3.097$, $p = .081$), but this variable interacted significantly with stop place ($F(1,134) = 9.592$, $p = .002$. Figure 13c). The figure shows that the difference between [k]’s compact bursts and [t]’s diffuse ones is larger after [s] than after [?].

*** Please put Figure 13 about here. ***

When we consider that the alveolar fricative [s] occurs before [u] and the palatoalveolar fricative [?] before [i] and then compare Figure 13c with Figure 13b, we can see that, when unconfounded from vowel backness, the fricatives have the opposite effect on the compactness of the burst’s spectrum from the vowels that determine their place of articulation.

To explain the effects of fricative place, we can again appeal to the location of the palatoalveolar constriction between [k]’s velar and [t]’s alveolar constrictions. By pulling the [t] constriction back, [?] 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 because the fricative is articulated as far forward as the stop, even if [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 the alveolar as the palatoalveolar fricative.

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] 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. This vowel 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 the vowels are opposite those of coarticulation with the fricatives, a result that raises the possibility that compensating for coarticulation with one context would cancel out the compensation for coarticulation with the other. No evidence for cancellation was observed in any listeners’ responses. 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.

4.4.4 Second formant frequency at the onset of voicing in the following [o]: In the analysis of F2 onset frequency in the contexts [ni, nu, ?i,? su?], vowel had a significant effect (*F*(1,281) = 152.42, $p < .001$) and voice did not (*F*(1,281) = 2.109, $p > .10$), but they interacted significantly with one another (*F*(1,281)=4.198, $p = .041$) and with stop (*F*(1,281) = 13.614, $p < .001$). This three-way interaction is displayed in Figure 14a, which shows that the effect of a voiceless vowel is greater than that of a voiced one on the velar stop [k], but the effect of a voiced vowel is larger than that of a voiceless one on the alveolar stop [t]. Nonetheless, vowel backness affects F2’s onset frequency in the same direction after both [k] and [t]: values are higher after [i] than [u].

In the analysis that unconfounds vowel backness from fricative place (contexts [ni, nu, ?i,??u?], vowel and voice are both significant (*F*(1,286) = 76.403, $p < .001$, *F*(1,286) = 17.248, *p* < .001), they interact (*F*(1,286) = 22.192, *p* < .001), and both interact with stop place (*F*(1,286) = 5.378, *p* = .021, Figure 14b). Figure 14b shows that the coarticulatory effect of the voiceless vowels is smaller than that of the voiced ones, especially for the velar stop [k]. Nonetheless, when the effects of vowel backness are analyzed separately for voiceless and voiced vowels, the effect of vowel and its interaction with stop place reached significance for voiceless as well as voiced vowels (voiceless: vowel *F*(1,136) = 5.928, *p* = .016; vowel by stop *F*(1,136) = 4.224, *p* = .042; compare voiced: vowel *F*(1,141) = 134.335, *p* < .001; vowel by stop *F*(1,141) = 30.391, *p* < .001).

In the analysis of the contexts [?u,?su?], which unconfounds fricative place from vowel backness, the effect of 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 14c). The figure shows that F2 onset frequency is uniformly lower after [s] than [?], even when the vowel intervening phonologically between the fricative and stop is held constant. The effects of the fricatives are essentially identical to those of the vowels that trigger the allophonic variation in fricative place: F2 is lower after [s] as well as [u] than after [?] as well as [i?].

*** Please put Figure 14 about here. ***

On the one hand, higher F2 onset frequencies are expected following [i] than [u] because the front vowel’s own F2 is much higher than the back vowel’s by virtue of its constriction at the front of the palate and spread lip aperture. On the other hand, it is at first surprising that F2 values are lower after [s] compared to [?], as [s] itself has a fronter articulation than [?] and its noise energy is concentrated at much higher frequencies. That higher frequency concentration of energy during the [s] itself is so high because it only depends on the length of the cavity in front of the constriction, which is considerably shorter in [s] than [?]. The constriction is too close for any back cavity resonance to radiate from the mouth. However, by the time F2’s onset frequency is measured at voice onset, the oral cavity had opened considerably and the back cavity resonances can radiate noticeably. Because the back cavity is absolutely long in both [s] and [?], F2 arises from it, and because that cavity is relatively longer for the fricative with the fronter articulation, [s], it produces a lower F2 than the one with the backer articulation, [?]. This finding is consistent with other studies that have compared F2 onset frequencies following [s] vs [?] (Mann & Repp, 1980; Whalen, 1981; Nittrouer, Studdert-Kennedy, & McGowan, 1989; Jongman, Wayland, & Wong, 2000).

If these acoustic effects determined how listeners compensated for coarticulation with the fricative and voiceless vowel contexts, then their stop place judgments would be biased in the same direction after the voiceless back vowel [u?as the alveolar fricative [s], that is, more “t” responses because coarticulation with both segments lowers F2 onset frequency compared to coarticulation with the voiceless front vowel [i?and the palatoalveolar fricative [?]. Like the analyses of center of gravity differences, this prediction is disconfirmed by the results of Experiment 1, where both English and Japanese listeners instead gave fewer “t” responses after [u?and [s] than after [i?and [?] .

4.4.5 Summary and discussion: This acoustic study examined three measures of the coarticulation of the stops [k] and [t] with preceding fricative and vowel contexts: the center of gravity and standard deviation of the stops' burst spectra and F2's frequency at the onset of the following [o]. The purpose of these examinations was to determine which preceding gestures might influence the stops' acoustics enough that listeners might compensate for those influences in identifying the stops' place. A preceding fricative's place of articulation significantly affected all three measures of coarticulation. The burst spectra's centers of gravity also clearly differed as 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 as much if not more after voiceless as voiced vowels, and F2's onset frequency likewise differed after voiceless as well as voiced vowels, even if not as much.

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. Two of these measures, the bursts' centers of gravity and F2's onset frequency, were altered in the same direction by both [u] and [s] and in the opposite direction by both [i] and [?]. Listeners might therefore be expected to compensate in the same way for coarticulation with the vowel as with the fricative in each pair. This parallelism is striking because the strident fricative is pronounced as alveolar [s] before the back vowel [u] and as palatoalveolar [?] before the front vowel [i]. This expectation is not fulfilled: English listeners respond “t” more often after voiced [u] than voiced [i] but less often after [s] than [?], and Japanese listeners also respond “t” less often after [s] than [?]. In light of the acoustic evidence showing that the voiceless vowels’ gestures are produced along with the fricatives’ gestures and that the following stops coarticulate with both gestures, it is especially striking that neither the English nor the Japanese listeners compensated for the vowels’ articulations. These findings thus disconfirm the predictions of Hypothesis 2a. The third measure of coarticulation, the standard deviation of the stop bursts’ spectra, differs from the other two in being altered in opposite directions by the [u] and [s] or by [i] and [?]. Because these vowels’ and fricatives’ coarticulatory effects alter the stops’ acoustics in opposite directions, compensating for one could cancel the effect of compensating for the other, as predicted by Hypothesis 2b. This prediction is disconfirmed by the absence of any evidence of cancellation. The only version of the compensation for coarticulation account that is not challenged by these results is the one in which listeners compensate for coarticulation with the fricative alone, Hypothesis 2c. In light of the extensive evidence just presented of the stops’ coarticulation with preceding voiceless as well as voiced vowels, it is surprising that listeners do not compensate for coarticulation with the vowel as well as the fricative. The compensation for coarticulation account must be amended in some fashion to permit listeners to compensate for just one coarticulated gesture and not all.

None of these acoustic findings challenge the auditory contrast account because it attributes stop place biases to the energy distribution across the spectrum in the preceding context rather than that context’s gestures and their coarticulatory effects. Because [su?] concentrates energy high in the spectrum, a following stop whose burst spectrum concentrates energy at intermediate frequencies will sound lower and thus more like [k], while after the low frequency energy concentration in [?i?], that same stop will sound higher and more like [t].

4.5 The absence of a vowel effect

There is still a puzzle here: why does a preceding voiced vowel not influence Japanese listeners’ judgments of the following stop, while it does influence English listeners’? Experiment 2 showed that this difference did not reflect a difference in the syllabic affiliation of the stop and preceding vowel between Japanese and English.

There is another possibility. As can be seen in the spectrograms in Figure 2, the differences in spectral energy concentrations between [s] and [?] are not in the same frequency range as those between [i] and [u]. The principal acoustic differences between the fricatives are all above 3000 Hz, while those between the vowels are all below this frequency. It is thus possible, perhaps even probable, that the spectral contrast between the following stops and the fricatives would differ from that between them and the vowels. There is even evidence for such a difference in the English listeners' responses, which are influenced by the preceding fricatives influence responses at the [t] end of the continuum but in the middle of the continuum by the preceding vowel contexts. Even so, we do not yet see how to exploit this difference between the fricative and vowel contexts to explain why the latter have no influence at all on Japanese listeners' stop place judgments. Experiments are being prepared to investigate this issue further.

4.6 The absence of affiliation effects

Experiment 2 showed that Japanese listeners' indifference to the preceding vowel could not be explained by its being syllabified separately from the stop, and Experiment 3 showed that a preceding fricative affected English listeners' stop place judgments just as much when it could syllabify with a preceding vowel as when there was no preceding vowel to syllabify with. The latter finding is a failure to replicate a result reported by Mann & Repp (1981) and may also cast doubt on Samuel & Pitt's (2003) argument that the fricative's effect varies inversely with how strongly it affiliates with the preceding string. These findings are also at odds with the conclusion of Dupoux and his colleagues (Dupoux, et al., 1999, 2001; Dehaene-Lambertz, et al., 2000) that a language's phonotactics separate segments into syllables at the earliest detectable stage of processing.

If segments were separated into syllables during perception, then those that belong to the same syllable should influence one another's perception more than those that belong to different syllables. There is evidence that segments are syllabified by the time certain phonotactic constraints apply or segments are separated into words. For example, Moreton (2002) shows that a bias against hearing [l] after [d] disappears when the [d] is preceded by a vowel with which it can be syllabified. Kirk (2001) demonstrates comprehensively that listeners syllabify segments prior to hypothesizing where words begin and end. However, the issue for us here is not *whether* listeners syllabify segments during perception but *when*. The absence of syllable affiliation effects in Experiments 2 and 3 indicates that the non-phonotactic, context-induced response biases that we have studied here do not depend on syllabification, but instead arise before segments have been grouped into syllables.

Syllabification is expected to have no effect if there is a prior, linguistically naïve stage of processing when only the auditory qualities of successive segments have any perceptual effect. If the percept of one segment's auditory qualities contrasts with that of its neighbor's during this stage, that contrast may induce a response bias later on. This hypothesized auditory stage of processing knows no more about syllabification than it does about any other linguistic fact. That in Experiment 3 the preceding fricatives affect English listeners' biases as much when a vowel precedes as when none does confirms this expectation. The failure to find any effect on Japanese listeners' biases of the syllabic affiliation of the stop and the preceding vowel in Experiment 2 also suggests that these context effects arise before segments are syllabified.

4.7 Concluding remarks

The experiments reported in this paper were designed to answer two questions. First, does a sound's context affect its perception by contrasting with it auditorily or because listeners compensate for coarticulation with that context? Second, do contextual effects arise before listeners apply their linguistic knowledge to the signal? The results of these experiments indicate that a sound's perception is altered by auditory contrast with its context and not via compensation for coarticulation with that context, and that this alteration occurs before listeners apply their linguistic knowledge. These results pose a new question in their turn: why does context sometimes have no effect? Future experiments seeks an answer to this question.

References

Amano, S., & Kondo, T. (2000). *NTT database series: Lexical properties of Japanese, 2^{nd} 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/

Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: A crosslinguistic 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*, 25, 1568-1578.

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

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes, *Journal of Memory and Language*, 27, 143-165.

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 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.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of Acoustical Society of America*, 108, 710-722.

Itô, J. (1986). *Syllable Theory in Prosodic Phonology*, Ph.D. dissertation, University of Massachusetts, Amherst.

Itô, J. (1989). A prosodic theory of epenthesis, *Natural Language and Linguistic* Theory, 7, 217-259.

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.

Kahn, D. (1976). *Syllable-based Generalizations in English Phonology*, Ph.D. dissertation, MIT.

Kingston, J. (2005). From ears to categories: New arguments for autonomy, S. Frota, M. Vigario, & M. J. Freitas, (eds.) *Prosodies: Proceedings of the First Conference on Phonetics and Phonology in Iberia (PaPI)*, (pp. 177-222), Berlin: Mouton de Gruyter.

Kirk, C. J. (2001). *Phonological Constraints on the Segmentation of Continuous Speech*, Ph.D. dissertation, University of Massachusetts, Amherst.

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. & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*), *Journal of Acoustical Society of America*, 113, 53-56.

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.

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

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on the perception of the [?]-[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 Acoustical Society of America*, 69, 548-558.

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

McCawley, J. D. (1977). Accent in Japanese, L. M. Hyman (ed.), *Studies in Stress and*

Accent, Los Angeles: University of Southern California, Department of Linguistics.

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

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters, *Cognition*, 84, 55-71.

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

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. (2000). Merging information in speech recognition: Feedback is never necessary, *Behavioral and Brain Sciences*, 23, 299-370.

Samuel, A.G., & Pitt, M.A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation, *Journal of Memory and Language*, 48, 416–434.

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.

Sugito, M ., Hirose, H. (1988). Production and perception of accented devoiced vowels in

Japanese, *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, 22, 19-36.

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

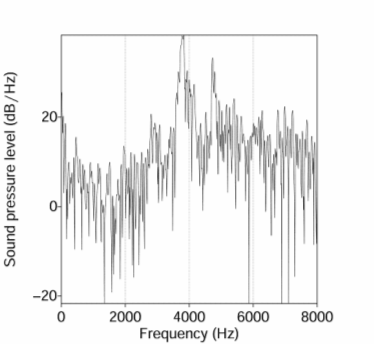
Tsuchida, A. (1997). *Phonetics and Phonology of Japanese Vowel Devoicing*, Ph.D. dissertation, Cornell University.

Vance, T. (2005). *An Introduction to Phonetics and Phonology for Teachers of Japanese*, Cambridge, UK: Cambridge University Press.

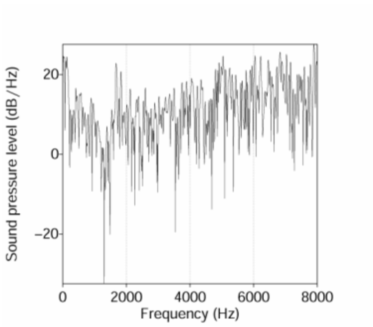
Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the /s-?/ boundary, *Journal of the Acoustical Society of America*, 69, 275-282.

(2b) Compensation for Coarticulation with the Voiceless Vowel and the Fricative						
Listener's Language	(1) Auditory Contrast or (2c) Compensation for Coarticulation with Fricative		(2a) Compensation for Coarticulation with the Voiceless Vowel		(3) Phonological Contra	
	Fricative	Stop Bias	Vowel	Stop Bias	Vowel or Fricative	Stop Bias
Japanese	(1) High [s] (2c) Front [s]	(1) Low [k] (2c) Back [k]	Back [u?	Front [t]	Low [u?	High [t]
	(1) Low [?] (2c) Back [?]	(1) High [t] (2c) Front [t]	Front [i?	Back [k]	High [i?	Low [k]
English	(1) High [s] (2c) Front [s]	(1) Low [k] (2c) Back [k]	Back [u?	Front [t]	High [s]	Low [k]
	(1) Low [?] (2c) Back [?]	(1) High [t] (2c) Front [t]	Front [i?	Back [k]	Low [?]	High [t]

Table I. Biases in Japanese and English listeners' stop place responses as a function of the preceding [su?vs [??contexts, as predicted by the (1) auditory contrast, (2a-c) compensation for coartculation, and (3) phonological contrast hypotheses. The left column in each pair identifies the relevant context, and the right the bias in stop place judgments. "High" and "low" refer to differences in energy concentrations in the contexts' spectra and to the perceived spectrum of the stop. "Front" and "back" refer to differences in the positions of articulators during the production of the contexts and their perceived positions in the stop.



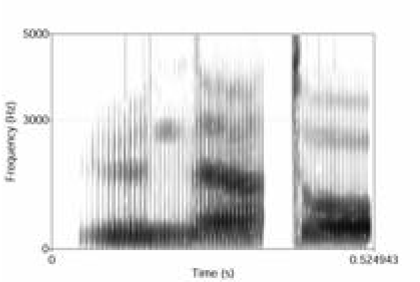
(a) [ʔiʔ]



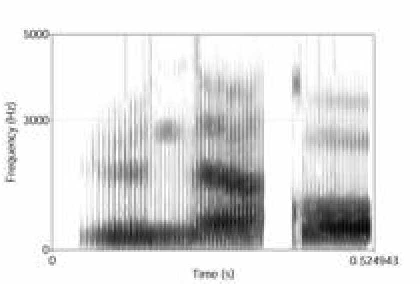
(b) [suʔ]

Figure 1: Spectra of (a) [ʔiʔ] and (b) [suʔ] showing their energy distributions between 0-8000 Hz. The spectra were calculated from a 50 ms wide Gaussian window centered in each fricative.

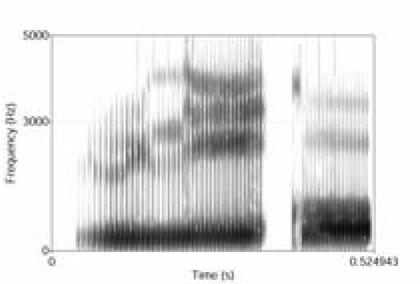
(a) [? u n a t o]



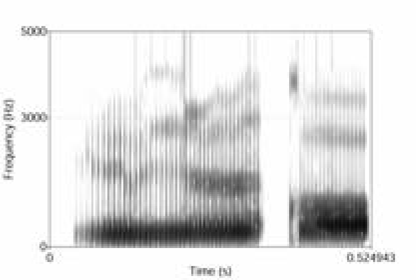
(b) [? u n a k o]



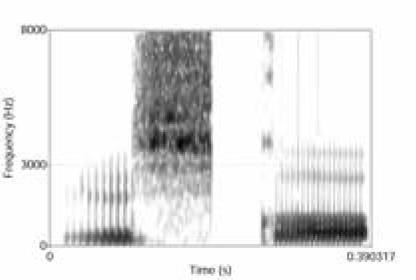
(c) [? u n i k o]



(d) [? u n u k o]



(e) [? u ?i? k o]



(f) [? u s u? k o]

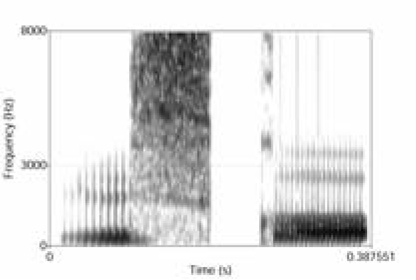


Figure 2: Spectrograms of example stimuli: (a) [?unato], (b) [?unako], (c) [?uniko], (d) [?unuko], (e) [?u?i?ko], and (f) [?usu?ko].

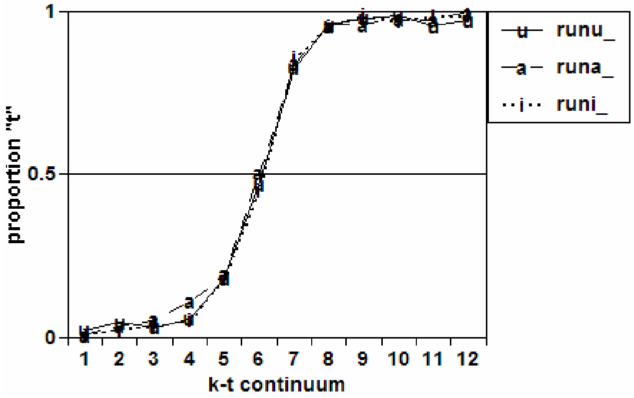


Figure 3a: Mean proportions of Japanese listeners' "to" responses to each step along the [k-t] continuum in the three voiced vowel contexts.

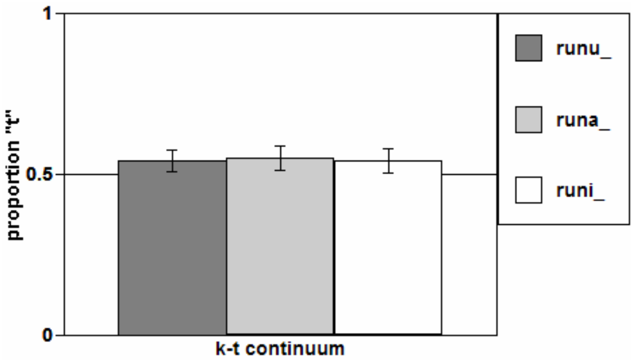


Figure 3b: Mean total proportions (95% confidence intervals) of Japanese listeners' "to" responses across the entire continuum in the three voiced vowel contexts.

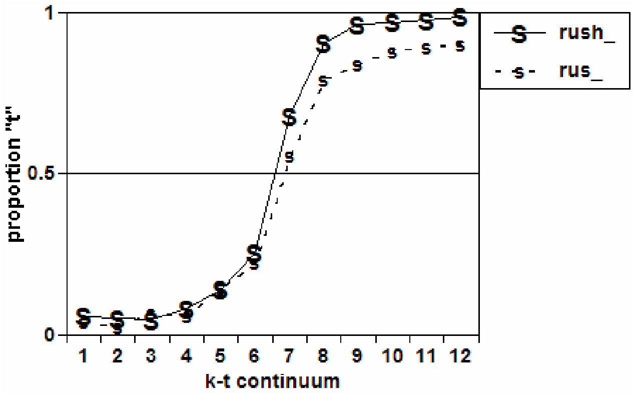


Figure 4a: Mean proportions of Japanese listeners’ “to” responses in the two fricative contexts: lower case “s” = [su], upper case “S” = [?i].

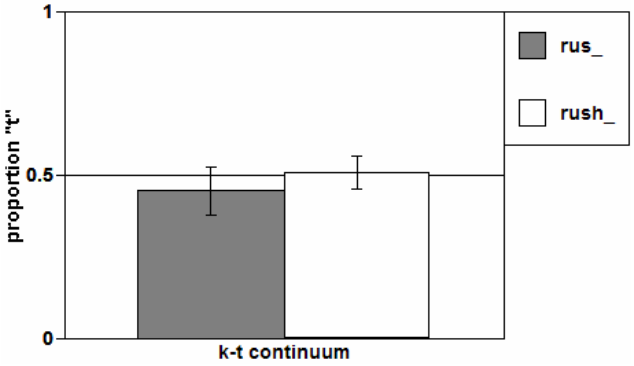


Figure 4b: Mean total proportions of Japanese listeners’ “to” responses in the two fricative contexts: lower case “s” = [su], upper case “S” = [?i].

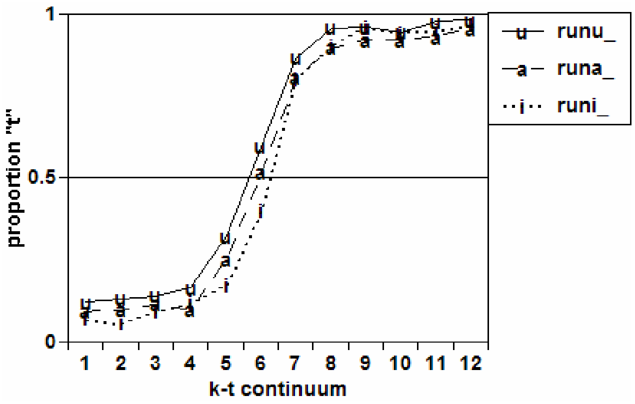


Figure 5a: Mean proportions of English listeners' "t" responses to each step along the [k-t] continuum in the three voiced vowel contexts.

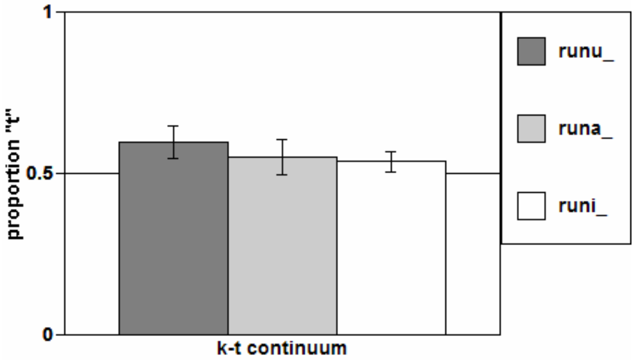


Figure 5b: Mean total proportions of English listeners' "t" responses in the three voiced vowel contexts.

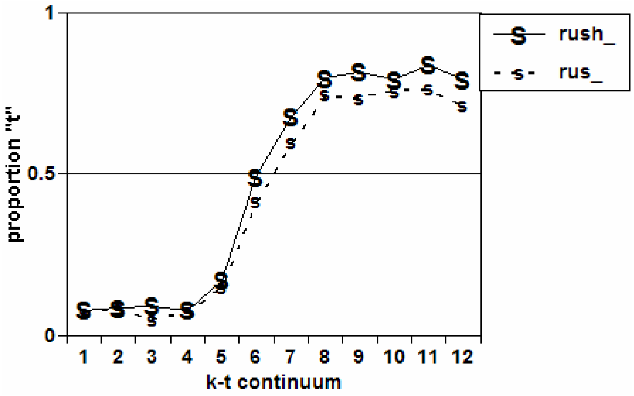


Figure 6a: Mean proportions of English listeners' "t" responses to each step along the [k-t] continuum in the two fricative contexts: lower case "s" = [su?], upper case "S" = [?i?].

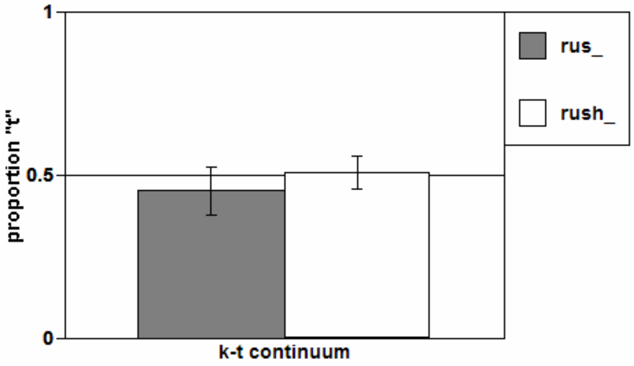


Figure 6b: Mean total proportions of English listeners' "t" responses in the two fricative contexts.

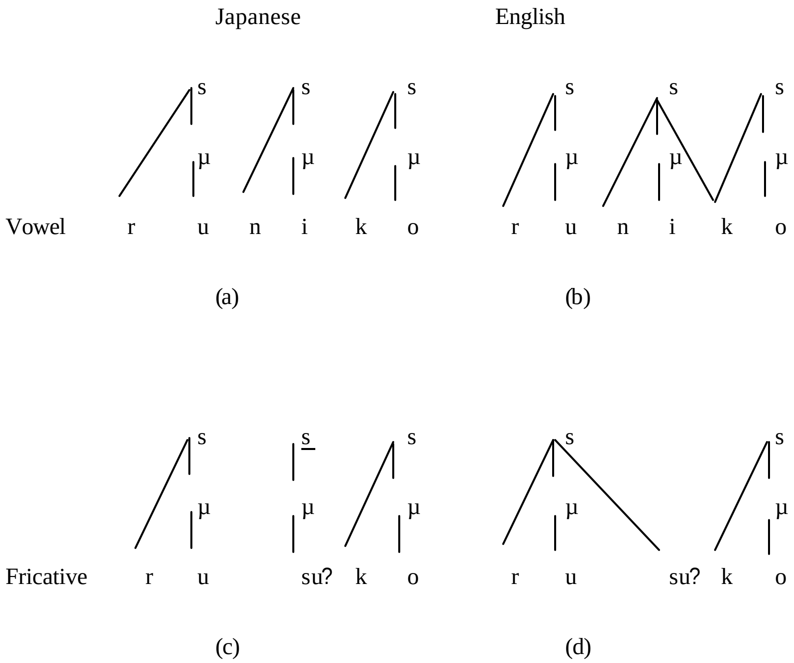


Figure 7: Predicted syllabifications of sequences containing (a,b) voiced and (c, d) voiceless vowels by (a, c) Japanese and (b, d) English listeners. “s” = syllable and “μ” = mora.

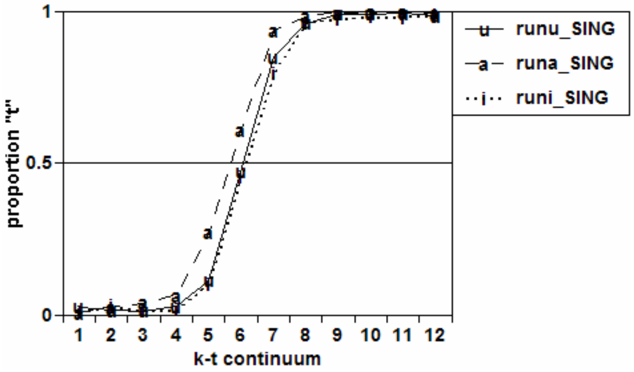


Figure 8a. Mean proportions of Japanese listeners' "t" responses to each step along the [k-t] continuum as a function of the quality of the preceding voiced vowel.

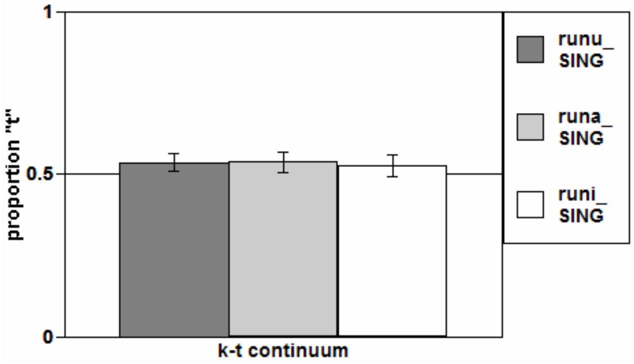


Figure 8b: Mean total proportions of Japanese listeners' "t" responses across the entire [k-t] continuum as a function of the quality of the preceding voiced vowel.

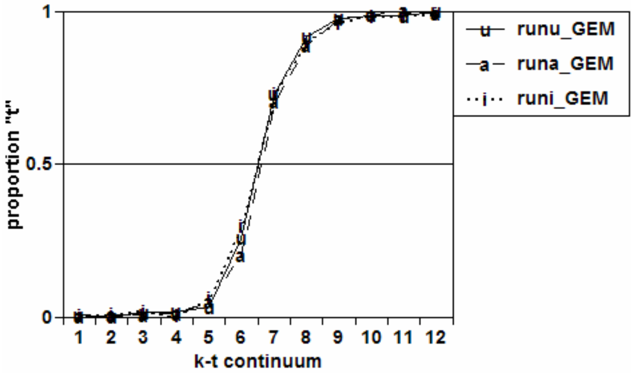


Figure 9a. Mean proportions of Japanese listeners' "tt" responses to each step along the [kk-tt] continuum as a function of the quality of the preceding voiced vowel.

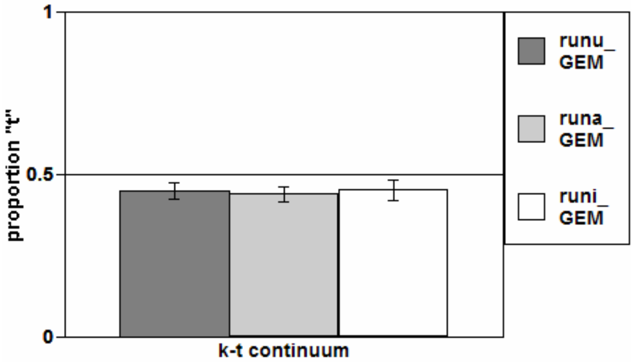


Figure 9b: Mean total proportions of Japanese listeners' "t" responses across the entire [k-t] continuum as a function of the quality of the preceding voiced vowel.

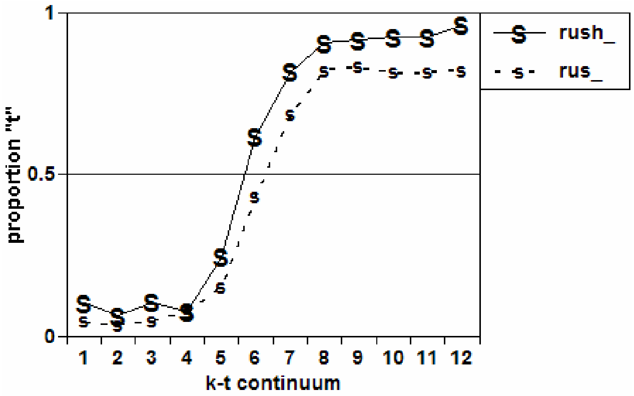


Figure 10a: Mean proportions of English listeners' "t" responses to each step in the [k-t] continuum as a function of the place of the preceding fricative, when it is preceded by another syllable: lower case "s" = [s], upper case "S" = [?].

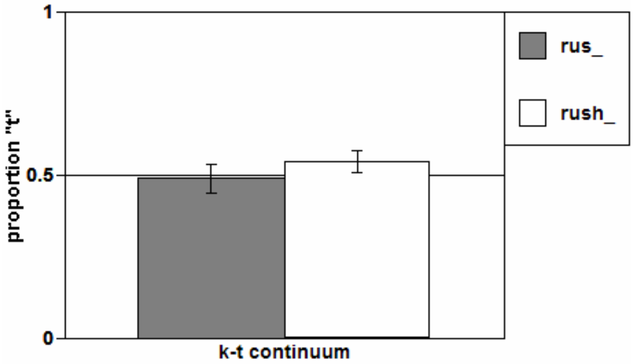


Figure 10b: Mean total proportions of English listeners' "t" responses across the entire [k-t] continuum as a function of the place of the preceding fricative, when it is preceded by another syllable.

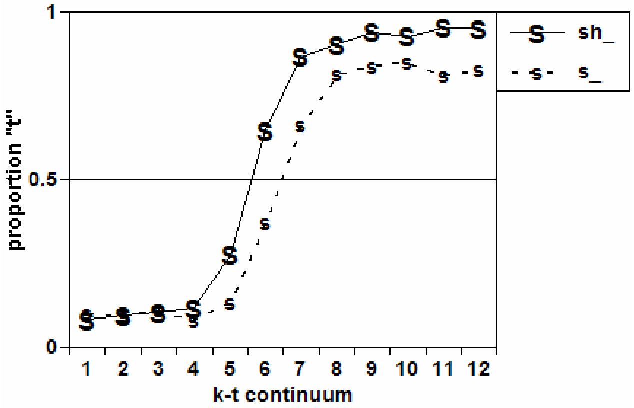


Figure 11a: Mean proportions of English listeners' "t" responses for each step in the [k-t] continuum as a function of the place of the preceding fricative, when it is not preceded by another syllable: lower case "s" = [s], upper case "S" = [?].

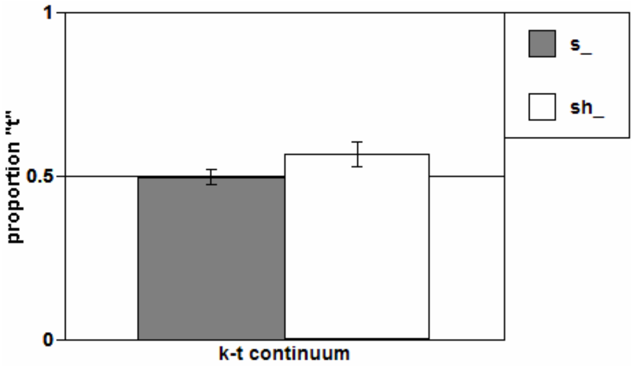


Figure 11b: Mean total proportions of English listeners "t" responses across the entire [k-t] continuum as a function of the place of the preceding fricative, when it is not preceded by another syllable, averaged across the k-t continuum.

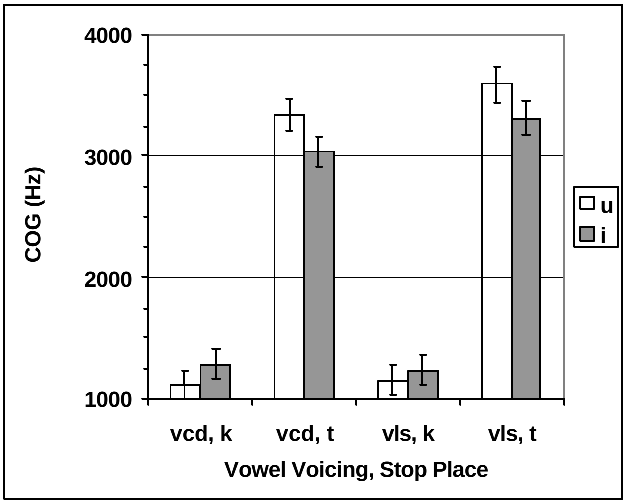


Figure 12a. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [u, i]. Voiceless [u?] and [i?] are preceded by and [s] [?], respectively.

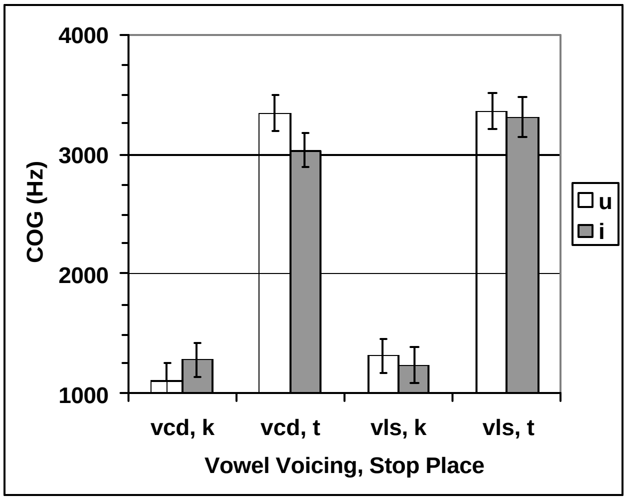


Figure 12b. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [?].

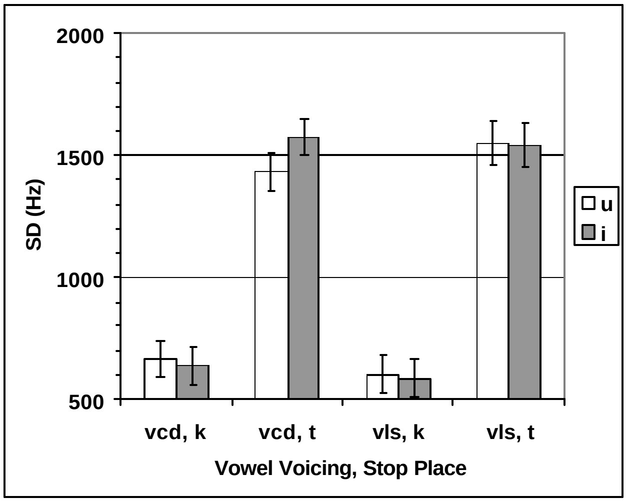


Figure 13a. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [u, i]. Voiceless [uʔ] and [iʔ] are simultaneous with [s] and [ʔ], respectively.

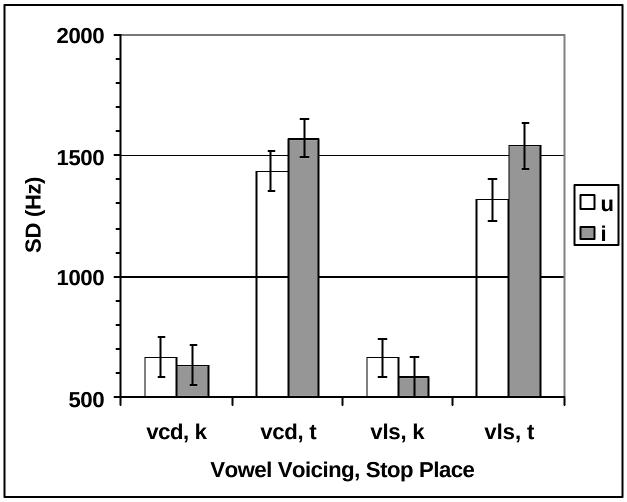


Figure 13b. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [ʔ].

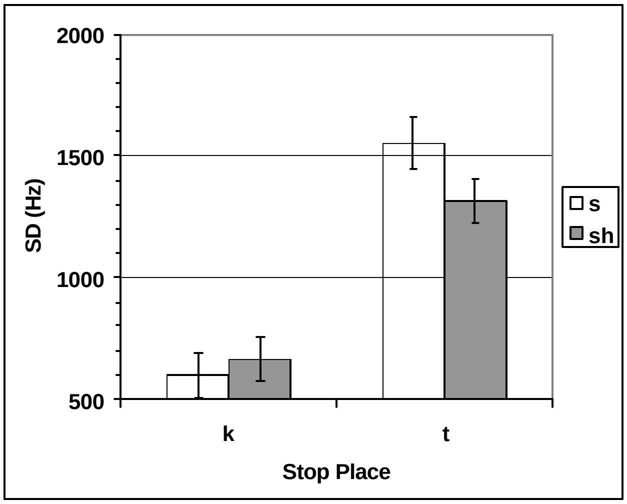


Figure 13c. Mean standard deviation values (95% confidence intervals) of [k] and [t] stop bursts, preceded by alveolar vs palatoalveolar fricatives, [s] vs [?]. Both voiceless vowels are [u?].

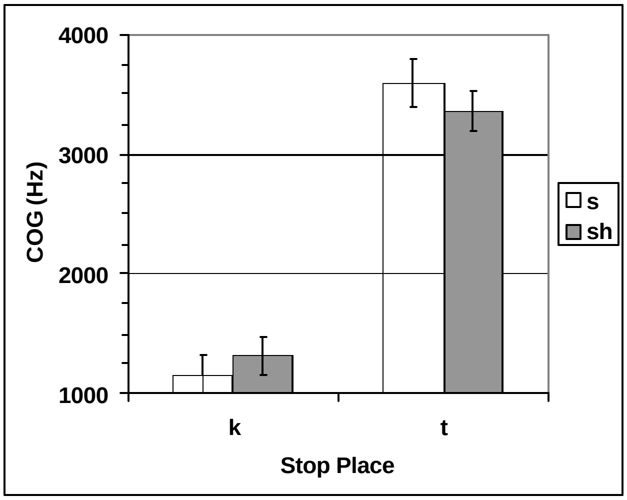


Figure 12c. Mean center of gravity values (95% confidence intervals) of [k] and [t] stop bursts, preceded by voiced and voiceless [u, i]. Voiceless [u?and [i?are simultaneous with [s] and [?], respectively.

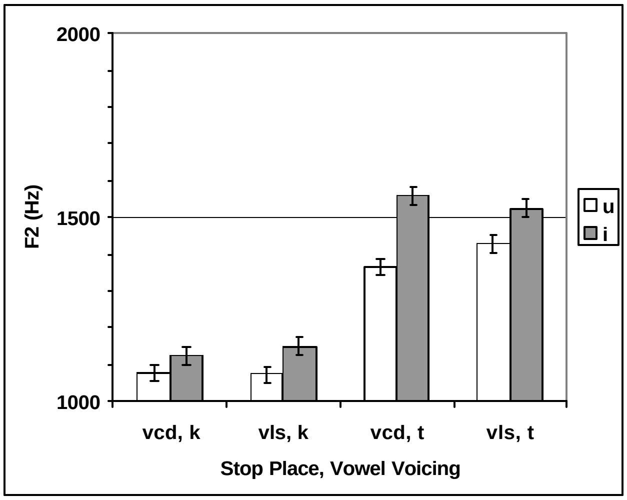


Figure 14a. Mean F2 values (95% confidence intervals) at voice onset following [k] vs [t], preceded by voiced and voiceless [u, i]. Voiceless [u? and [i? are simultaneous with [s] and [?], respectively.

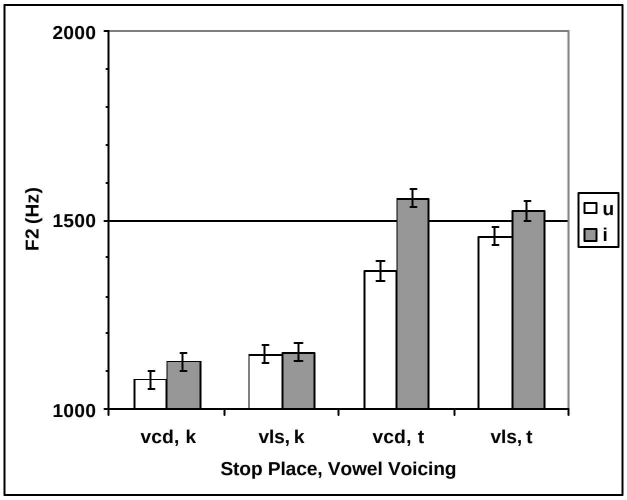


Figure 14b. Mean F2 values (95% confidence intervals) at voice onset following [k] vs [t], preceded by voiced and voiceless [i, u]. Both voiceless vowels are simultaneous with [?].

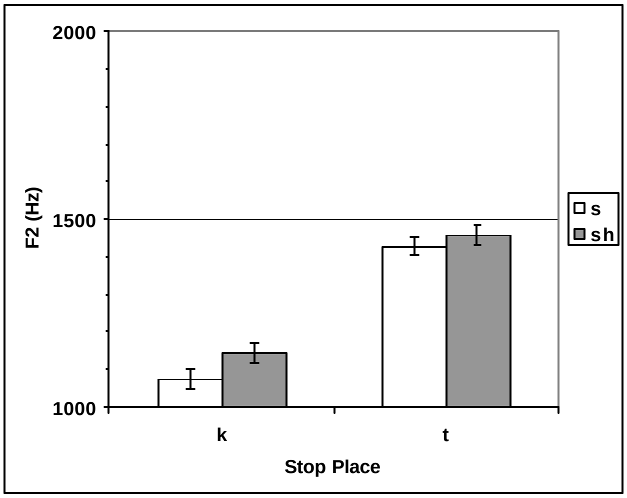


Figure 14c. Mean F2 values (95% confidence intervals) at voice onset following [k] vs [t], preceded by alveolar vs palatoalveolar fricatives, [s] vs [?]. Both voiceless vowels are [u?.