

**Processing Missing Vowels:  
Allophonic Processing in Japanese**

Naomi Ogasawara\*, \*\* & Natasha Warner\*, †  
\* Department of Linguistics, University of Arizona  
\*\* Department of English, National Taiwan Normal University  
† Max Planck Institute for Psycholinguistics, Nijmegen

Short title: Processing Japanese Vowel Reduction

Corresponding author:  
Natasha Warner

Address: Department of Linguistics  
University of Arizona  
P.O. Box 210028  
Tucson, AZ 85721-0028  
U. S. A.

Until 7/08: Max Planck Institute for Psycholinguistics  
Box 310  
6500 AH Nijmegen  
The Netherlands

Phone: +31-24-352-1382  
FAX: +31-24-352-1213  
Email: naomi703@ntnu.edu.tw (Naomi Ogasawara)  
nwarner@u.arizona.edu (Natasha Warner)

### Abstract

The acoustic realization of a speech sound varies, often showing allophonic variation triggered by surrounding sounds. Listeners recognize words and sounds well despite such variation, and even make use of allophonic variability in processing. This study reports five experiments on processing of the reduced/unreduced allophonic alternation of Japanese high vowels. The results show that listeners use phonological knowledge of their native language during phoneme processing and word recognition. However, interactions of the phonological and acoustic effects differ in these two processes. A facilitatory phonological effect and an inhibitory acoustic effect cancel one another out in phoneme processing; while in word recognition, the facilitatory phonological effect overrides the inhibitory acoustic effect. Four potential models of the processing of allophonic variation are discussed. The results can be accommodated in two of them, but require additional assumptions or modifications to the models, and primarily support lexical specification of allophonic variability. (147 words)

### Acknowledgements

We are grateful to James McQueen, Holger Mitterer, Andy Wedel, Tim Vance, Adam Ussishkin, Merrill Garrett, Haruo Kubozono, and Erin Good for discussion on this material, and to the Cognitive Science Program and the Department of Linguistics at University of Arizona for support for travel to Japan. We would also like to thank Ikuo Hara, Akira Watanabe, Tomoko Yoshino, Miyuki Takasawa, Kyoko Masuda, Keiichi Tajima, Takayuki Arai, and Kazuki Kuwabara for their help in making the experiments in Japan possible. Any errors or misinterpretations are our own.

## Introduction

The /k/s in the English words *kit* and *skit* are acoustically rather different, yet listeners identify both sounds as /k/ and successfully recognize words containing both types of /k/. In Japanese, allophonic variation leads to reduction or even deletion of the high vowels /i, u/ when they occur between two voiceless consonants (Figure 1). The first vowel in the word [k(i)ta]<sup>1</sup> 'North' often consists only of palatalized voiceless noise at the release of the [k], while a full [i] vowel is present in [kinoo] 'yesterday.' When listeners hear acoustically different allophones, such as the two kinds of English /k/ or reduced vs. full Japanese /i/, do they map the allophones onto a single, more abstract phonemic category? Do listeners take into account the conditioning environment that determines which allophone will occur? Do such non-distinctive, low-level alternations affect recognition of words as well as of sounds? What mechanism in the spoken word recognition or speech perception system is responsible for processing allophonic variability?

(FIGURE 1 ABOUT HERE)

The alternation in Japanese vowels is traditionally referred to as vowel devoicing, rather than reduction. The high vowels /i, u/ are expected to be devoiced between voiceless consonants (e.g., *kita* [k(i)ta] 'North') or between a voiceless consonant and the end of the utterance (e.g., *aki* [ak(i)] 'autumn') (Vance, 1987). We refer to these as "devoicing environments." Phonetically, so-called devoiced vowels are often deleted (Vance, 1987, in press; Yuen, 2000), or a short, low-amplitude vowel may remain (Yuen, 2000). We refer to all such vowels as "reduced" to avoid the debate over devoicing vs. deletion.<sup>2</sup> Japanese vowels, both unreduced and reduced, cause coarticulation in the preceding consonant, which allows identification of the /i/ or /u/ even if the

vowel itself is deleted (Ostreicher & Sharf, 1976). The reduced vowels are considered to be present at least in the underlying form, because they cause coarticulation, and because, if they were not present at all, this would leave consonant clusters (e.g. /kt/ in [k(i)ta] ‘North’) that are otherwise phonotactically impossible in Japanese. We investigate how listeners process these reduced vowels at the sound level and during spoken word recognition.

Allophonic variability is rampant in the world’s languages. Listeners’ allophonic processing has been most extensively investigated for assimilation across word boundaries in English or Dutch, as in “garde[m] bench” for “garden bench” (Gaskell, Hare, & Marslen-Wilson, 1995; Gaskell & Marslen-Wilson, 1998; Gow, 2001, 2002; Mitterer & Blomert, 2003). Mitterer, Csépe, and Blomert (2006) extend this work to manner assimilation in Hungarian, and Gow and Im (2004) examine across-word boundary assimilation of voicing or place in Hungarian and Korean. Widely varying tasks (phoneme monitoring, lexical decision, priming, and even EEG measurement of mismatch negativity) have generally shown inhibited processing when allophones are placed in inappropriate environments (e.g. “garde[m]” before a velar rather than a labial consonant). Results have been somewhat mixed on whether the effect applies equally in words and non-words or more strongly in words. The main question for this line of research is how listeners map a sound that is acoustically a good match to one phoneme (e.g. /m/) onto a different phoneme (e.g. /n/), and do so only when this is phonologically appropriate. This question of mapping one phoneme onto another is different from the situation in the current study, though. In the case of vowel reduction, the two allophones, [(i)] and [i], correspond only to a single phoneme /i/. The question is how listeners process a single sound with more than one realization, rather than how listeners convert one sound into another, and processing is within a word. However, the issue of appropriateness of environment is the same.

This literature on processing of across-word-boundary assimilation shows that native listeners of a language are able to process sounds that occur as various allophones, and that listeners take the phonological environment and language-specific allophonic alternations into account when doing so. There are several possibilities for how to model this. Mitterer and colleagues (Mitterer & Blomert, 2003; Mitterer et al., 2006) suggest that compensation for coarticulation across word boundaries occurs as part of low-level auditory processing, based partly on a finding of an MMN (EEG mismatch negativity) response to inappropriate allophones at a very early time point in the signal. Gaskell et al. (1995) and Gaskell and Marslen-Wilson (1998) discuss a higher level phonological inference mechanism, in which the listener uses knowledge of phonological patterns of the language to infer the underlying form from the allophone. This might work by listeners recognizing a context after the allophone and regressively deducing the identity of the allophone (e.g. recognizing /b/ in “garde[m] bench” and inferring that the [m] is underlyingly /n/). Mitterer and McQueen (submitted) find evidence for such a mechanism for processing of reduced or deleted word-final consonants. Gow (2001) argues that cues to the following segment are present in the assimilated segment, instead.

A third possibility is that the lexicon contains information about allophones rather than underlying segments (so that “garde[m]” in fact is one of the underlying representations of “garden”). Spinelli, McQueen, and Cutler (2003) discuss such a solution for an alternation involving French [R]. Finally, an exemplar model of speech perception (e.g. Goldinger, 1998) provides a fourth possible solution, in which all differences in sounds would be represented in the lexicon, because acoustic traces of previously heard exemplars would be stored. Mitterer and McQueen (submitted) review these models and some additional variations on them, dividing them into lexical (allomorph or exemplar listings) and prelexical (auditory or phonological

inference) methods. The current study provides a test of these four ways to explain allophonic processing, and does so for a different language and a very different type of alternation than the extensively studied across-word-boundary assimilation.

Although the phonetics of Japanese reduced vowels has been well studied (as discussed above), the processing of them has not. Dupoux, Hirose, Pallier and Mehler (1999) and Dupoux, Pallier, Kakehi, and Mehler (2001), however, find that Japanese listeners are more likely than French listeners to hear an /u/ in sequences like /eb(u)zo/ even if the /u/ is completely absent (where /bz/ would be an impossible cluster in Japanese). This shows listeners' sensitivity to phonotactic patterns. However, their stimuli include both devoicing and voicing environments (e.g. /eb(u)zo/ with voiced consonants), so the implications for processing of vowel reduction are not clear. Cutler, Otake, and McQueen (submitted) investigate listeners' recognition of words embedded next to consonants in Japanese, e.g. recognition of /sake/ 'salmon' in a sequence such as /nyaksake/. This sequence is only possible in Japanese if the listener assumes there to be a reduced /u/, making the form /nyakusake/. Cutler et al. (submitted) find that listeners do not restore reduced vowels at an early, automatic stage of processing. Furthermore, they find that the number of words in the lexicon containing a given string with a reduced vowel affects how likely listeners are to assume a reduced vowel is present. They conclude that Japanese listeners' restoration of reduced vowels happens during lexical, rather than prelexical, processing.

Using phoneme monitoring and lexical decision tasks, we investigate processing of reduced vowels themselves, rather than their surrounding environment. Reduced vowels are acoustically weak, which might make them harder to process. However, phonotactic knowledge (which indicates that a vowel must be present because of the consonant cluster) should facilitate the recognition even of reduced vowels. Moreover, language-specific knowledge of the

allophonic alternation should facilitate recognition of reduced and unreduced vowels in their appropriate environments (e.g., [(i)] in [k(i)ta] ‘North’ and [i] in [itʃiɡo] ‘strawberry’).

The question is how these three effects (acoustic strength, phonotactic knowledge, and allophonic appropriateness) interact. If the influence of allophonic appropriateness is stronger than other effects, for example, then reduced vowels should be easier to process in the appropriate environment than unreduced vowels, despite their acoustic weakness. Five experiments were conducted: Experiments 1 and 3 examined detection of the vowel /i/ by Japanese listeners, using phoneme monitoring. Experiment 2 tested American English listeners, who lack Japanese phonological knowledge. Experiments 4 and 5 explored Japanese listeners’ ability to recognize words containing allophonic variation, using a lexical decision task. By comparing the results from the two tasks, we can see whether the effects of acoustic cues and phonological knowledge are consistent in the processing of both sounds and words. This allows a new test of the four models of allophonic processing described above.

### Experiment 1

This experiment tests processing of reduced and unreduced /i/ in appropriate and inappropriate environments. Japanese listeners monitored for the phoneme /i/ (realized as either full [i] or reduced [(i)]) in nonsense words in three phonological environments: the devoicing environment (between two voiceless consonants, e.g. /hokito/), the voicing environment (after the voiced consonant [dʒ], e.g. /tadʒiga<sup>3</sup>), and the nasal environment (after /n/, e.g. /kedanida/). If knowledge of allophonic alternations affects processing, each allophone should be easier to recognize in its appropriate environment. However, reduced [(i)] has weaker acoustic cues than unreduced [i], which might make [(i)] more difficult overall. Furthermore, reduced vowels have

shorter durations than unreduced vowels, since reduced vowels in this study have no voiced part associated with the vowel. Cutler, van Ooijen, Norris, and Sánchez-Casas (1996) have shown that RTs in phoneme monitoring correlate inversely with vowel duration, which would also lead to faster RTs for unreduced vowels. Thus, the relative strength of the allophonic effect vs. the acoustic strength/duration effect can be judged, particularly in the devoicing environment, where these effects conflict: allophonic knowledge favors the appropriate, reduced, vowel, but acoustic strength/duration favors the unreduced vowel. In the voicing environment (e.g., /tadʒiga/), both effects favor the unreduced vowel (it is appropriate and has stronger acoustic cues).

The nasal condition, although it is also a voicing environment because of the voiced nasal consonant, separates the effects of listeners' allophonic vs. phonotactic knowledge. NC (Nasal-Consonant) clusters are phonotactically legal in Japanese (e.g. /Nt/ in /haNtai/ 'opposite'), whereas clusters of two non-nasal consonants are not (e.g. \*/hakdai/). Therefore, listeners are not forced to identify an /i/ when they hear a sequence such as [kedap(i)da] (perhaps with deleted /i/), since /kedaNda/ is phonotactically possible. In this condition, we predict that the disadvantage for the reduced vowel will be even greater than in other conditions, because listeners receive no help from phonotactics in identifying the reduced vowel.

### *Methods*

*Materials.* This experiment tests two factors (Table 1): reduction of the vowel /i/ (reduced vs. unreduced), and phonological environment (devoicing, voicing, and nasal). Reduced [(i)] is the appropriate allophone in the devoicing environment, at least for careful speech in the Tokyo dialect. Unreduced [i] is the appropriate allophone in the voicing and nasal environments.

(TABLE 1 ABOUT HERE)

Thirty test items, phonotactically legal Japanese non-words with 3 or 4 moras, were created for each environment (Appendix A). Each item was realized once with an unreduced [i] and once with a reduced [(i)]. The phoneme /i/ appeared only once in each item, in the penultimate mora (CVCiCV or CVCVCiCV). There were 30 fillers containing an unreduced [i] in other positions, 300 fillers without /i/, and 10 similar practice items, all with 2-4 moras. Syllable structures in the fillers and practice items were varied, containing some geminates, coda nasals, and long vowels, to make the materials more word-like for Japanese.

The first author, a female native speaker of the Tokyo dialect, recorded the stimuli in a recording booth using a high quality microphone and a CD recorder sampling at 44.1 kHz. All items were produced naturally without any subsequent editing of the recording. Waveforms and spectrograms of all experimental items were examined to confirm that formants and a periodic wave were visible for the [i] in the unreduced conditions, but not for [(i)] in the reduced conditions. Appendix A provides information about the acoustics of the stimuli.

*Participants.* Forty nine native speakers of the Tokyo dialect participated. They were students at universities in Tokyo or acquaintances of the first author, 18-42 years old, without speech or hearing disorders. They received a small gift or payment for participating. All but four also took part in Experiment 4 before Experiment 1, with a short break in between. Experiment 4 (lexical decision) was ordered first to avoid training listeners to focus on a particular phoneme before doing the more general lexical decision task.

*Procedures.* Items were counterbalanced for the reduction factor and placed in two counterbalanced lists with the fillers. Each list had a different order of stimuli with at least two

non-/i/ fillers between items containing /i/. The experiment took place in quiet locations such as classrooms or libraries in five universities in the Tokyo area, and in private homes. The E-Prime software (Psychology Software Tools, Inc.), running on a laptop, controlled the experiment, and stimuli were presented over headphones. Participants were instructed orally and in writing to press a button on the response box as quickly as possible when they heard a syllable which contained the vowel /i/, as in /i, ki, ʃi, tʃi, ni, hi, mi, ri, gi, dʒi, bi, pi/. This instruction was necessary because of the *kana* syllabary system in Japanese: if participants were instructed to respond to /i/, they would only respond to the onsetless V syllable /i/, not to CV syllables containing /i/. Participants were presented with the practice test and then with one of the two experimental lists. Afterwards, participants filled out a language background questionnaire and read some Japanese sentences aloud, but this speech was not recorded. The experimenter confirmed auditorily that participants produced reduced vowels in devoicing environments.

### *Results*

RTs were measured from the offset of the /i/. For unreduced [i], as in [hok*i*to], this was the end of voicing, or the end of the second formant if the following consonant was voiced. For reduced [(i)], as in [hok(*i*)to], the end of [(i)] was taken to occur at the end of the preceding consonant and onset of the following consonant (which are the same point), since there was no portion of the signal uniquely associated with /i/ and distinguishable from the surrounding consonants.

Acoustic cues to [(i)] may be located in both the preceding and following consonants, but cues to unreduced [i] can also spread into neighboring consonants. Any RTs outside of 200-1500 ms were treated as errors, which excluded 8.4% of the data. Three subjects' data were excluded from the by-subjects analysis because they failed to respond to any items in one condition.

Analyses of variance were carried out on the RTs and error rates, by subjects ( $F_1$ ) and by items ( $F_2$ ). Each ANOVA had the factors environment (devoicing, voicing, and nasal), reduction (reduced and unreduced), and (for subjects analyses only) counterbalanced group. In the subjects analyses, environment and reduction were within-subjects factors and counterbalanced group was a between-subjects factor. In the items analyses, environment was a between-items factor and reduction was a within-items factor. The Greenhouse-Geisser correction was applied whenever the sphericity assumption was violated.

Results appear in Figure 2. For RTs, the main effects of environment ( $F_1(2, 88) = 6.3, p < .005; F_2(2, 87) = 6.3, p < .005$ ) and reduction ( $F_1(1, 44) = 19.1, p < .001; F_2(1, 87) = 11.1, p < .005$ ) were significant, as was their interaction ( $F_1(2, 88) = 4.2, p < .05; F_2(2, 87) = 3.4, p < .05$ ). Therefore, the simple effects of reduction were tested. In the devoicing environment, reduction had no significant effect ( $F_1$  and  $F_2 < 1$ ), but in the voicing environment, RTs were faster to unreduced [i] than to reduced [(i)] (significant by subjects and nearly by items,  $F_1(1, 44) = 18.4, p < .001; F_2(1, 29) = 4.1, p = .052$ ). The nasal environment also showed slower RTs for reduced [(i)] ( $F_1(1, 44) = 13.4, p < .005; F_2(1, 29) = 9.2, p < .01$ ). Thus, reduced vowels were detected slowly in both environments where they were inappropriate, but where they were appropriate, they were detected as quickly as unreduced vowels. The pattern of results was the same even when RTs were measured from the onset instead of the end of the target mora, which demonstrates that the results do not depend on the choice of measurement time point.

(FIGURE 2A AND 2B ABOUT HERE)

For error rates (Figure 2b), the main effects of environment ( $F_1(2, 88) = 43.1, p < .001$ ;  $F_2(2, 87) = 42.3, p < .001$ ) and reduction ( $F_1(1, 44) = 54.7, p < .001$ ;  $F_2(1, 87) = 26.5, p < .001$ ) and their interaction ( $F_1(2, 88) = 51.1, p < .001$ ;  $F_2(2, 87) = 26.6, p < .001$ ) were all significant. Tests of the simple effect of reduction showed no effect in the devoicing environment ( $F_1$  and  $F_2 < 1$ ), or, unlike the RT data, in the voicing environment ( $F_1$  and  $F_2 < 1$ ). Only the nasal environment showed a significant effect of reduction ( $F_1(1, 44) = 75.9, p < .001$ ;  $F_2(1, 29) = 33.8, p < .001$ ), with performance less accurate, but still above 50% correct, for reduced [(i)]. The fact that listeners still responded relatively often to reduced [(i)] suggests that the stimuli did sound like possible Japanese pronunciations of the /ni/-consonant sequence.

### *Discussion*

Three phenomena might affect detection of the vowel: 1) acoustic cues are stronger and durations longer for unreduced vowels, which could lead to faster responses for unreduced vowels in all environments (Cutler et al., 1996). 2) Appropriateness of the allophone should favor the unreduced vowel in the voicing and nasal environments and the reduced vowel in the devoicing environment. 3) Phonotactic constraints should lead listeners to detect a vowel in the devoicing and voicing environments, but should not have any influence in the nasal environment, where phonotactic constraints do not require a vowel. The results show that Japanese listeners process both allophones equally quickly and accurately in the devoicing environment: /i/ in [hokito] and [hok(i)to] is recognized equally well. In the voicing environment, listeners detect reduced and unreduced vowels equally accurately, but they detect the reduced vowels more slowly. That is, /i/ in [tadʒiga] is recognized more quickly, but no more accurately, than /i/ in

[tad̥(i)ga]. Finally, in the nasal environment, listeners are both slower and far less accurate to detect reduced vowels (i.e. [ked̥(i)da] is more difficult than [ked̥iida].)

For the nasal environment, this result is straightforward: with weak acoustic cues, no phonotactic support for presence of a vowel, and an inappropriate allophone, reduced [(i)] in strings like [ked̥(i)da] is very difficult to detect. The voicing environment, in comparison, shows no effect of reduction on error rates, but does show one for RTs. We propose that listeners' phonotactic knowledge forces them to recognize a vowel in the reduction condition (e.g. in [tad̥(i)ga]), despite weak acoustic cues and allophonic inappropriateness, but these factors do slow processing. Finally, we propose that the lack of an effect of reduction in the devoicing environment (e.g. [hok(i)to]) reflects a cancelling out of effects: phonotactic knowledge forces recognition of a vowel, and allophonic appropriateness favors the reduced vowel. However, the weak acoustic cues and shorter duration for the reduced vowel inhibit detection, cancelling out the facilitation one might otherwise see. Because this represents a null effect in the devoicing environment, this interpretation will be revisited based on comparison with Experiment 4.

Experiment 1 indicates that Japanese listeners use language-specific knowledge of both allophonic and phonotactic patterns to process vowels, and this knowledge interacts with the effect of acoustic strength and vowel duration. To confirm the role of language-specific phonological knowledge, we replicated the experiment with English-speaking listeners.

## **Experiment 2**

We replicated Experiment 1 with American English listeners who had no knowledge of Japanese. Because stimuli are non-words, listeners with no knowledge of Japanese can do the task. English

does not have high-vowel reduction or a phonotactic constraint against consonant clusters like Japanese, so English listeners should not display effects of Japanese phonological patterns.

English does have deletion of /ə/, as in *parade* ([p<sup>h</sup>əréd] → [p<sup>h</sup>réd]) (Hammond, 1999; Patterson, LoCasto, & Connine, 2003), but since it is schwa that deletes, this will not lead English listeners to interpret a deleted vowel as /i/. They are only likely to respond based on acoustic cues.

Therefore, slower RTs and higher error rates for reduced [(i)] are predicted in all environments.

### *Method*

*Participants.* Forty-five native speakers of American English with no reported hearing problems were recruited from introductory linguistics courses at the University of Arizona. None of the participants had any knowledge of Japanese. Most had experience studying a foreign language, but none were fluent bilinguals. They received a small amount of course credit.

*Materials and Procedures.* Materials and procedures were the same as in Experiment 1 except that the instructions and the questionnaire were given in English. The task was conducted in a phonetics lab at the University of Arizona, in a quiet room but not in a sound attenuated booth, so as to create a testing environment similar to that of Experiment 1.

### *Results*

The data were analyzed in the same way as in Experiment 1. Responses outside 200-2000 ms, constituting 8.1% of the data, were treated as errors. This range is wider than in Experiment 1 because the task was more difficult, since the listeners were hearing stimuli pronounced in an unfamiliar language. 14 subjects failed to respond to any items in at least one reduced [(i)] condition. Seven items also elicited no response in the reduced condition. Because of the low

response rate, for RTs, only the averages rather than a statistical analysis are presented.

ANOVAs on the error rates are more reliable for this experiment. The subjects and items without responses are included in the graphs and in the error rate statistical analysis because their lack of responses is meaningful.

Figure 3 shows the RTs and error rates for the American listeners. Reduced [(i)] was detected more slowly than unreduced [i] in all environments. For errors, the main effects of environment ( $F_1(2, 86) = 18.1, p < .001$ ;  $F_2(2, 87) = 9.7, p < .001$ ) and reduction ( $F_1(1, 43) = 362.1, p < .001$ ;  $F_2(1, 87) = 527.0, p < .001$ ) were significant, with a significant interaction ( $F_1(2, 86) = 25.1, p < .001$ ;  $F_2(2, 87) = 10.1, p < .001$ ). Reduced [(i)] was recognized significantly less accurately in each environment (devoicing:  $F_1(1, 43) = 137.6, p < .001$ ;  $F_2(1, 29) = 74.1, p < .001$ ); voicing:  $F_1(1, 43) = 186.2, p < .001$ ;  $F_2(1, 29) = 233.1, p < .001$ ; nasal:  $F_1(1, 43) = 451.2, p < .001$ ;  $F_2(1, 29) = 301.8, p < .001$ ). The significant interaction, however, indicates that this effect is smaller for the devoicing environment than for the other two.

(FIGURE 3A AND 3B ABOUT HERE)

### *Discussion*

In general, English listeners had higher error rates than Japanese listeners. The American English subjects missed even unreduced vowels more than 20% of the time, while the mean error rates of the Japanese subjects never surpassed 15% even in the most difficult environment. This might be simply because American English subjects were not familiar with sounds in Japanese; therefore, they had more difficulty isolating the vowel from the surrounding sounds. Alternatively, Japanese /i/ might not be a perfect match to the English /i/ category.

The high error rates of American English listeners for reduced [(i)] in all environments confirm that they were not influenced by Japanese allophonic or phonotactic patterns, and confirm that [(i)] is harder to detect than [i] based on acoustic cues alone. Interestingly, despite the lack of phonological knowledge, the effect of reduction is largest for the nasal condition and smallest for the devoicing condition. This may reflect the strength of acoustic cues for reduced vowels in the various environments. In the nasal environment, the only acoustic cue to [(i)] is palatalization of the preceding nasal and perhaps of the following consonant, which can be very difficult to perceive. In the devoicing environment, the consonant before [(i)] is often a voiceless stop or affricate, so that the release noise provides strong coarticulatory information. The voicing environment is intermediate in acoustic strength of coarticulation. Thus, the pattern of results for American listeners mirrors the likely availability of acoustic cues.

### Experiment 3

Experiment 3 extends Experiment 1 in two ways: it broadens the range of environments tested in the voicing condition, and it utilizes cross-splicing. In Experiment 1, all voicing environment stimuli had the consonant /dʒ/ before the target vowel, because this is the only voiced non-nasal consonant which the speaker found easy to produce before inappropriately reduced [(i)]. To rule out the possibility that there is something particular to the /dʒ-/ environment, Experiment 3 used some stimuli which had a voiced consonant *after* the target vowel, instead. This allows for a more general test of the voicing environment, but it also introduces a new aspect to what kind of processing is tested: when listeners hear /dʒ/ before a vowel (e.g. /tadʒ.../), they already have enough information to know a reduced vowel is inappropriate. When they hear a voiceless

consonant before a vowel (e.g. /kek.../), they cannot rule out the possibility of reduction until they hear the following consonant (e.g. /kekiz.../). Thus, in Experiment 3, the voicing environment contains some stimuli that provide information about appropriateness of the target before it occurs, and others that only do so afterwards.

Furthermore, in Experiment 1, there could be systematic acoustic differences between the reduced and unreduced vowel stimuli other than the intended difference in vowel reduction. Therefore, in Experiment 3, the mora containing the target /i/ (i.e. the Ci sequence) was spliced between the two reduction conditions to create the stimuli, so that there were no acoustic differences between reduction conditions outside that mora.

### *Methods*

*Materials.* The same six (2 x 3) conditions as in Experiment 1 were created in the current experiment. In the voicing environment 10 items had /dʒ/ before /i/ (e.g. /tadʒiga/) and 10 had a voiced consonant after /i/ (e.g. /kekizo/) (Appendix B). Half of the stimuli in each environment were created by splicing the mora containing the target vowel from the reduced recording into the unreduced recording, and half were created by splicing that mora from the unreduced recording into the reduced recording. That is, in each environment, half of the stimuli were an unmanipulated recording for the unreduced condition but spliced for the reduced condition, and the other half were the reverse. The mora containing the target /i/ (e.g. /dʒi, ki/) was used for splicing, rather than just the vowel, because the preceding consonant is, in these materials, longer when a vowel is reduced (cf. Han, 1994), and this is likely to be a perceptual cue to vowel reduction. Materials were otherwise similar to those in Experiment 1.

*Participants.* Forty-seven university students similar to those in Experiment 1 were recruited. All but one participated in Experiment 5 (lexical decision) prior to this experiment.

*Procedures.* Instructions and procedures were identical to those in Experiment 1.

### *Results*

RTs were measured from the onset of the following consonant, as in Experiment 1. Any RTs outside the range between 200 ms to 1800 ms were treated as errors, which excluded 5.9% of the data. One subject's data were removed from the by-subjects analyses due to failure to respond in the nasal condition. ANOVAs used the same design as in Experiment 1. For RTs (Figure 4a), the main effect of reduction was significant ( $F_1(1, 44) = 6.40, p < .05; F_2(1, 56) = 4.50, p < .05$ ); but that of environment was not ( $F_1(2, 88) = 1.07, p > .1; F_2 < 1$ ). The interaction was significant by-subjects only ( $F_1(2, 88) = 3.75, p < .05; F_2(2, 56) = 1.42, p > .1$ ). Because of the partially significant interaction, the effect of reduction was tested for each environment. It was significant only for the nasal environment (devoicing:  $F_1(1, 44) = 2.61, p > .1, F_2(1, 19) = 1.18, p > .1$ ; voicing:  $F_1$  and  $F_2 < 1$ ; nasal:  $F_1(1, 44) = 7.63, p < .005; F_2(1, 18) = 3.27, p > .05$ ), with the reduced vowel detected more slowly. In the error rate analysis (Figure 4b), the main effects of environment ( $F_1(2, 88) = 39.61, p < .001; F_2(2, 56) = 21.08, p < .001$ ) and reduction ( $F_1(1, 44) = 29.1, p < .001; F_2(1, 56) = 13.89, p < .001$ ) and their interaction ( $F_1(2, 88) = 47.03, p < .001; F_2(2, 56) = 17.64, p < .001$ ) were all significant. The effect of reduction was significant only in the nasal environment (devoicing:  $F_1$  and  $F_2 < 1$ ; voicing:  $F_1$  and  $F_2 < 1$ ; nasal:  $F_1(1, 44) = 59.83, p < .001; F_2(1, 18) = 19.03, p < .001$ ).

(FIGURE 4A AND 4B ABOUT HERE)

These results differ from Experiment 1 in showing no effect of reduction even on RTs for the voicing environment. Because this might stem from the division of voiced environment stimuli into those with a preceding /dʒ/ (e.g. /tadʒiga/) and those with a following voiced consonant (e.g. /kekizo/), RTs were examined for these two sub-environments (Figure 5). Post-hoc ANOVAs were carried out with type of voicing environment (preceding vs. following voiced consonant) and the reduction of the target vowel as factors, for just the voicing environment data. No significant effects were found (preceding consonant:  $F_1$  and  $F_2 < 1$ ; reduction:  $F_1$  and  $F_2 < 1$ ; interaction: ( $F_1(1, 44) = 1.20, p > .1$ ;  $F_2(1, 18) = 1.27, p > .1$ ). However, there was a trend toward slower detection of reduced [(i)] only when it followed the voiced [dʒ].

(FIGURE 5 ABOUT HERE)

### *Discussion*

The results of this experiment confirm that the effects in Experiment 1 do not stem from uncontrolled acoustic differences between reduced and unreduced vowel stimuli in some other part of the item. With the exception of RTs for the voicing environment, the same pattern of effects holds as in Experiment 1. For the voicing environment, the crucial difference between the two experiments is the division of voicing environment stimuli into items with a voiced consonant before vs. after the target vowel (e.g. /tadʒiga/ vs. /kekizo/). When voiced /dʒ/ occurs before the target, listeners already know by the time they detect [(i)] that it is allophonically inappropriate. However, when the following consonant is what makes the environment a voicing

environment, listeners may detect the [i] before they have enough information about the following consonant to realize that vowel reduction is inappropriate. Although the difference between the two types of voicing environment items is not significant, this is probably because the number of items in this post hoc comparison is too small for sufficient statistical power. The trend in reaction times is for items with a preceding voiced consonant to show slower reaction times for reduced vowels, as in Experiment 1. Weber (2001) also discusses the issue of information about appropriateness of an allophone occurring before or after a target.

Experiments 1-3 have addressed effects of acoustic, phonological, and phonotactic knowledge on processing of sounds through the phoneme monitoring task. With Experiments 4-5, we turn to word-level processing.

#### **Experiment 4**

Experiment 4 extends the above findings by using a lexical decision task. The stimuli and conditions are similar to those in the earlier experiments, except that /i/ occurs in real words, e.g. /yak**is**oba/ 'fried noodles' for the devoicing environment, /od**z**isan/ 'uncle' for the voicing environment, and /butan**ik**u/ 'pork' for the nasal environment.

#### *Method*

*Participants.* Forty-seven native listeners of the Tokyo dialect participated. All but two subsequently participated in Experiment 1.

*Materials.* The conditions parallel those for Experiment 1, but all target items were 3-6 mora real words. Twenty words were chosen for each environment (Appendix C), and each was recorded with reduced and unreduced /i/ as for Experiment 1. The vowel /i/ appeared only once,

in word-medial position, in most target items. Where /i/ occurred more than once (e.g. /onigiri/ 'rice ball,' as was necessary to find enough items), only the target /i/ (the /i/ following the nasal in /onigiri/) was varied for the reduction conditions. All reduced vowels in all environments were phonetically deleted, with coarticulation with neighboring consonants. In the nasal environment, where /i/ could, phonotactically, be deleted, deletion of /i/ from the items does not create an alternative real word. For example, if the /i/ in /ganimata/ 'bowlegged' is missed, there is no real word /gaNmata/ that could lead to a correct lexical decision response through recognition of the wrong word. There were 30 real-word fillers (some containing unreduced [i]), 200 non-word fillers (some containing appropriate reduced vowels), and 10 similar practice items.

*Procedures.* Procedures were the same as in Experiment 1 except for the instructions. Participants were informed that they would hear real and nonsense words, and were asked to press a button on a response box as quickly as possible when they heard a real word.

### *Results*

RTs were measured from the end of the word, and were excluded if faster than 50 ms or slower than 800 ms (7.6% of the data). One item in the voicing environment was excluded because it failed to elicit a response from any subject. Two items in the voicing environment were excluded because their target vowels were later found to contain some voicing .

ANOVAs were carried out as in Experiment 1. RTs (Figure 6a) showed significant effects of environment ( $F_1(2, 90) = 56.52, p < .001; F_2(2, 54) = 6.57, p < .005$ ), reduction ( $F_1(1, 45) = 49.39, p < .001; F_2(1, 54) = 10.02, p < .005$ ), and their interaction ( $F_1(2, 90) = 102.32, p < .001; F_2(2, 54) = 18.51, p < .001$ ). In the devoicing environment, RTs were faster for reduced than unreduced forms ( $F_1(1, 45) = 50.1, p < .001; F_2(1, 19) = 11.3, p < .01$ ). That is, [yak(i)soba]

'fried noodles' is easier to recognize than [yakisoba], even with the /i/ phonetically deleted. In the other environments, the reverse was true (voicing:  $F_1(1, 45) = 47.07, p < .001$ ;  $F_2(1, 16) = 4.57, p < .05$ ; nasal:  $F_1(1, 45) = 123.1, p < .001$ ;  $F_2(1, 19) = 31.1, p < .001$ ). The effect of reduction was greater in the nasal than in the voicing environment. The results were similar for error rates (Figure 6b) (environment:  $F_1(2, 90) = 48.95, p < .001$ ;  $F_2(2, 54) = 11.05, p < .001$ ; reduction:  $F_1(1, 45) = 57.27, p < .001$ ;  $F_2(1, 54) = 18.72, p < .001$ ; interaction:  $F_1(2, 90) = 56.65, p < .001$ ;  $F_2(2, 54) = 21.96, p < .001$ ). In the devoicing environment, subjects were more accurate for words with a reduced [i] ( $F_1(1, 45) = 19.27, p < .001$ ;  $F_2(1, 19) = 3.55, p = .075$ ). The opposite was true in voicing and nasal environments (voicing:  $F_1(1, 45) = 14.92, p < .001$ ;  $F_2(1, 16) = 8.10, p < .02$ ; nasal:  $F_1(1, 45) = 78.9, p < .001$ ;  $F_2(1, 19) = 34.47, p < .001$ ).

(FIGURE 6A AND 6B ABOUT HERE)

### *Discussion*

In all environments, use of the appropriate allophone led to faster word recognition, even when the appropriate allophone had the weaker acoustic cues. This indicates that, for word recognition, the facilitatory effect of phonological appropriateness not only canceled out but even surpassed any inhibitory effect of weaker cues for reduced vowels. Words like /yakisoba/ were easier to recognize with the underlying /i/ reduced than with it fully present. This differs from the phoneme monitoring results (Experiment 1), where phonological appropriateness and weak acoustic cues cancel each other out in the devoicing environment, leaving no effect of reduction.

One additional difference is that Experiment 1 (phoneme monitoring) showed faster, but not more accurate, processing of the appropriate allophone in the voicing environment (e.g.

/tadziga/). In Experiment 4, reduction in the voicing environment affected both RTs and errors. For Experiment 1, we argued that the phonotactic constraint against consonant clusters forced listeners to perceive the reduced [(i)] in the voicing environment, albeit slowly. This suggests that phonotactic knowledge plays a greater role in processing sounds, as in phoneme monitoring, than it does in processing words in lexical decision. When monitoring for a particular sound, it is not surprising if listeners make strong use of phonotactic knowledge to help them determine where that sound is likely to occur. When trying to decide simply whether a whole stimulus is a real word or not, however, they may not realize exactly what is strange about a word containing just one inappropriate allophone. Thus, phonotactic knowledge may allow listeners to avoid errors, if slowly, at the sound level but not at the word level.

As with Experiment 1 above, there could be other systematic acoustic differences between reduced and unreduced vowel conditions. Furthermore, the voicing environment stimuli for Experiment 4 also all had the consonant /dʒ/ before the /i/, creating a restricted set of environments. Experiment 5 addresses these issues.

### **Experiment 5**

Experiment 5 parallels Experiment 3 above, for the lexical decision task.

#### *Method*

*Materials.* The materials (Appendix D) were as in Experiment 4, with the same modifications applied as between Experiments 1 and 3 above (splicing and replacement of half the voicing environment stimuli with pre-voiced-consonant stimuli). Thus, half the voicing environment stimuli had a voiced consonant after /i/, as in /sekidome/ 'cough syrup.'

*Participants.* The participants were 46 of the 47 participants from Experiment 3.

*Procedures.* The procedure was the same as in Experiment 4.

### *Results*

RTs (Figure 7a) were measured from the end of the word. RTs outside the range between 0-900 ms were treated as errors, which excluded 4.9% of the data. One subject's data were removed from the by-subjects analyses due to failure to respond in one condition. ANOVAs confirmed that the main effects of environment ( $F_1(2, 86) = 32.12, p < .001; F_2(2, 57) = 4.39, p < .05$ ) and reduction ( $F_1(1, 43) = 49.96, p < .001; F_2(1, 57) = 22.51, p < .001$ ) and their interaction ( $F_1(2, 86) = 67.82, p < .001; F_2(2, 57) = 24.22, p < .001$ ) were all significant. Each environment showed a significant effect of reduction, with faster responses for words containing the appropriate allophone in all conditions (devoicing:  $F_1(1, 43) = 32.06, p < .001; F_2(1, 19) = 6.52, p < .05$ ; voicing:  $F_1(1, 43) = 27.49, p < .001; F_2(1, 19) = 10.92, p < .005$ ; nasal  $F_1(1, 43) = 93.13, p < .001; F_2(1, 19) = 40.88, p < .001$ ). For error rates (Figure 7b), the main effects of environment ( $F_1(2, 86) = 76.93, p < .001; F_2(2, 57) = 14.42, p < .001$ ) and reduction ( $F_1(1, 43) = 68.81, p < .001; F_2(1, 57) = 19.07, p < .001$ ) and their interaction ( $F_1(2, 86) = 47.02, p < .001; F_2(2, 57) = 15.97, p < .001$ ) were all significant. The effect of reduction was significant in each environment (devoicing:  $F_1(1, 43) = 15.71, p < .001, F_2(1, 19) = 9.85, p < .05$ ; voicing:  $F_1(1, 43) = 18.55, p < .001, F_2(1, 19) = 6.88, p < .05$ ; nasal:  $F_1(1, 43) = 72.69, p < .001; F_2(1, 19) = 20.46, p < .001$ ), and largely paralleled the RT patterns.

(FIGURE 7A AND 7B ABOUT HERE)

As in Experiment 3, RTs for the voicing environment were divided (Figure 8) by whether the voiced consonant preceded or followed the /i/ (e.g. /odʒisan/ 'uncle' vs. /sekidome/ 'cough syrup'). Post-hoc ANOVAs were carried out for just the voicing environment. The effect of the voicing environment type was significant by subjects ( $F_1(1, 43) = 7.43, p < .01; F_2(1, 18) = 1.07, p > .1$ ), with pre-voiced-consonant items (e.g. /sekidome/) showing slightly slower RTs. The main effect of the reduction was significant, showing slower recognition for words with reduced vowels ( $F_1(1, 43) = 22.44, p < .001; F_2(1, 18) = 10.44, p < .01$ ), but the interaction was not significant ( $F_1$  and  $F_2 < 1$ ). Unlike in Experiment 3, for lexical decision, listeners were slower to respond to words with inappropriately reduced vowels whether the information about inappropriate allophonic environment appeared before or after the vowel itself.

(FIGURE 8 ABOUT HERE)

### *Discussion*

The results of Experiment 5 were similar to those of Experiment 4, even for both types of voicing environment stimuli separately. The similar findings despite the use of splicing confirm that the results in Experiment 4 derive from the mora containing the reduced/unreduced vowel, not from uncontrolled acoustic differences elsewhere. Furthermore, when the two sub-types of voicing environment are examined, both show similar effects of reduction. Thus, when listeners are processing words for lexical decision, they react to inappropriate allophones in the same way regardless of when the information about inappropriateness becomes available. When listeners are monitoring for phonemes, however, their reaction is influenced by whether information about inappropriateness comes before or after the allophone itself. This difference reflects the time

course of processing words vs. sounds in these tasks. Since the /i/ is near the middle of each stimulus, for lexical decision, listeners must hear both the preceding and following consonant as well as other sounds after that before they can recognize the word. When monitoring for /i/, though, listeners can begin to respond as soon as they have enough acoustic information to detect an /i/, which may be before they hear much information about the following consonant.

### **General Discussion**

#### *The Interaction of Acoustic Cues and Knowledge of Allophonic Variation*

In lexical decision, the results showed quicker and more accurate processing of words containing the appropriate allophone, even when that allophone was reduced and thus had weak acoustic cues. That is, listeners found words easier to recognize when allophonic patterns of the language were followed. However, in phoneme monitoring, listeners found the appropriate allophone easier to detect only when the unreduced allophone was the appropriate one. In the devoicing environment, they found both allophones equally easy to detect.

We believe this shows that in processing of both words and sounds, there is an effect of appropriateness of allophones. However, in processing of sounds (as for phoneme monitoring on non-words), listeners pay more attention to acoustic cues than they do during lexical processing. In all environments, unreduced [i] has stronger acoustic cues than reduced [(i)]. Unreduced vowels also have greater duration, and previous research (Cutler et al., 1996) shows that vowels with longer duration have faster RTs. Therefore, in the phoneme monitoring experiment, the devoicing environment shows no difference between allophones because the appropriateness effect (in favor of [(i)]) cancels out the acoustic cue strength and duration effect (in favor of [i]).

*Differentiating Two Kinds of Phonological Knowledge*

Many previous studies have shown that language-specific phonological knowledge affects processing (McQueen, 2007). The current study separates two kinds of phonological knowledge: knowledge of allophonic variation, and knowledge of phonotactic patterns. Several comparisons show this distinction. First, in the basic phoneme monitoring experiment (Experiment 1), Japanese listeners respond differently to the voicing environment vs. the nasal environment. In the nasal environment, where phonotactic knowledge does not force recognition of a vowel (e.g. [kedaŋ(i)da] could be interpreted as /kedaNda/), listeners are both inaccurate and slow to detect reduced [(i)]. However, in the voicing environment, where phonotactic knowledge forces recognition of a vowel (e.g. [tadʒ(i)ga] cannot be interpreted as illegal \*/tadʒga/), they are slow but accurate in detecting it. Because the reduced vowel is inappropriate for both of these conditions, allophonic knowledge about reduction could not lead to this result.

Second, comparing the voicing and the devoicing environments, for both tasks (Experiments 1 and 4) it is clear that appropriateness of allophonic variation also affects processing. Although the devoicing environment does not show an effect of reduction in Experiment 1, it does in Experiment 4, and even in Experiment 1 this environment clearly elicits different results from the voicing environment. Both the voicing and devoicing environments share the phonotactic property of requiring a vowel (e.g. \*/hokto/, \*/tadʒga/), so phonotactic knowledge cannot account for this effect. Third, comparing Japanese and American listeners (Experiments 1 and 2) further demonstrates that language-specific phonological knowledge is involved, because American listeners show a different pattern. Thus, the current results show that

listeners can make use of two separate types of language-specific phonological knowledge in processing speech: knowledge of both allophonic alternations and phonotactic constraints.

*Better Performance for a Non-underlying Form*

Previous studies of processing of allophonic variation that used across-word-boundary place assimilation (e.g. “garde[m] bench” for “garden bench”) show that allophonic variation in an inappropriate environment hinders processing, and that listeners can use allophones to predict upcoming sounds in some cases (e.g. Gow & Im, 2004), but they do not necessarily show that appropriately realized allophonic variation actually facilitates processing of the allophone itself (Gaskell et al., 1995; Gaskell & Marslen-Wilson, 1998; Gow, 2001, 2002; Mitterer & Blomert, 2003). Similarly, Ranbom and Connine (2007) show that although listeners are tolerant of appropriate reductions of [nt] to nasal flap (e.g. “ge[nt]le” vs. a production that rhymes with “kennel”), they still find it easier to process the unreduced, underlying form. There is even some evidence of appropriate allophonic change hindering recognition (Gaskell & Marslen-Wilson, 1998). The current study shows that a reduced, non-underlying form can actually be easier to process than a clear, underlying form. This may be because across-word-boundary assimilation and nasal flapping are optional, so that the underlying, unaltered form (e.g. “garde[n] bench, ge[nt]le”) is always a possible pronunciation.

Although there is variability in Japanese vowel reduction (Kitahara, 1988; Vance, 1987, in press), it is much closer to obligatory than across-word assimilation or nasal flapping is in English. Hence, a reduced vowel is the only appropriate realization in the devoicing environment, at least for the Tokyo dialect. [kita] ‘North’ with an unreduced vowel is simply wrong, unlike “garde[n] bench.” The current study shows particularly in Experiments 4 and 5 (lexical decision)

that listeners actually recognize words with an appropriately reduced vowel more easily than words with an inappropriate but unreduced vowel. This is notable both because the reduced vowels are not the underlying form, and because they have weaker acoustic cues than the unreduced vowels. A form further from the underlying representation can be easier to process if it is the most appropriate form. This is a very different result from the findings on cross-word place assimilation, and it suggests a stronger role for phonological knowledge in environment-dependent processing than has previously been demonstrated.

### *The Mechanism for Processing Allophonic Alternations*

Four potential ways to model the processing of allophonic alternations were introduced above:

1) an auditory processing account that attributes allophonic effects to a very early, automatic processing stage, 2) a phonological inference mechanism that allows listeners to apply abstract knowledge to infer which allophones are appropriate, 3) putting allophonic information into lexical representations, and 4) an exemplar model relying on reference to past exemplars.

All of these except the phonological inference approach could model the appropriateness effect found in the current lexical decision data (Experiments 4 and 5). An auditory processing account might suggest that listeners' processing of speech is slowed when they hear unfamiliar sequences containing inappropriate allophones, as in [kit] or [dʒ(i)g]. A lexical representation account would involve the lexical representation containing the allophone [(i)], where appropriate, rather than the phoneme /i/: the lexical representation of /kita/ 'North' would be [k(i)ta], with no abstraction across [(i)] and [i]. The input [k(i)ta] would activate the lexical representation of the word for 'North' more strongly than [kita] would, since the latter would not

entirely match the lexical representation. An exemplar model would suggest that appropriate allophones in real words are recognized more easily than inappropriate allophones simply because listeners have heard the words far more times with the appropriate allophone, so the perceptual system contains many exemplars of the words with the appropriate allophones (e.g. [k(i)ta]) and only a few with the inappropriate allophones (e.g. [kita]). The phonological inference method, however, cannot account for the reduced, non-underlying vowel leading to easier word recognition (in the appropriate environment) than productions containing the unreduced, underlying vowel. The phonological inference model is about how listeners infer what the underlying segment is, so the underlying, unaltered segment should always be at least as easy to process as the non-underlying variant. That is, an allophonic variant should never lead to *improved* recognition in the phonological inference model, as we find.

Turning to the phoneme monitoring results, the remaining three models cannot all explain the results of Experiments 1 and 3. In phoneme monitoring, as in lexical decision, there was a difference between the devoicing and the voicing environments, suggesting that appropriateness of allophones does matter. However, in the devoicing environment, there was no difference in responses to the appropriate and inappropriate allophones, which we interpret as showing that the unreduced allophone is easier to recognize overall because of its short duration (Cutler et al., 1996) and stronger acoustic cues. There are thus three things any model needs to account for: 1) facilitation of appropriate allophones even in non-words, 2) slower processing of allophones with weaker acoustic cues, and 3) limitation of this penalty for weak acoustic cues to processing of non-words or to a phoneme monitoring task.

The auditory processing mechanism can easily include an allophonic appropriateness effect (1), along with a separate effect of overall difficulty in detecting reduced vowels (2), even

in non-word strings. However, it could not limit the acoustic weakness effect to non-word or sound-level processing (3). Auditory processing refers solely to processing of strings of sounds, so it should apply in the same way in non-words and real words, and regardless of whether the task involves sounds or words. Because the auditory processing account posits that allophonic processing takes place early and automatically, it should apply regardless of task. The phonological inference method, already ruled out by the lexical decision results, would have the same problem with the phoneme monitoring results as auditory processing does, because it also refers only to strings of sounds regardless of lexical status or task. The phoneme monitoring data also pose a further problem for the auditory processing account: low-level auditory processing cannot account for English listeners' failure to show appropriateness effects (Experiment 2).

The lexical representation mechanism would need a modification in order to explain the phoneme monitoring results, since representing words allophonically in the lexicon would not influence processing in non-words. One could potentially accommodate the phoneme monitoring results in the Merge Model (Norris, McQueen, & Cutler, 2000), for example, by including both [i] and [(i)] as categories in the phoneme decision module of Merge, so that responses to the two allophones could differ even in non-words. The appropriateness effect in non-words (1 above) could stem from comparison to real words that have material in common with the non-words. For example, if a listener hears the non-word string [hok(i)to], the string [k(i)t] would partially activate the real word [k(i)ta] 'North.' The activation from this word would then spread activation to the [(i)] category in the phoneme decision module (thus actually an allophone decision module) of Merge. The preference for appropriate allophones at the lexical level would thus spread to similar non-word strings. The only further assumptions necessary in Merge are that allophone categories with weak acoustic cues are decided upon more slowly (2 above), and

that this only affects categories in the phoneme decision module, not word-level categories at the lexical level (3 above), or that this is specific to the phoneme-monitoring task. Thus, the Merge Model can account for the current results, if lexical representations contain allophonic information (as proposed by Spinelli, et al. (2003)), if the phoneme decision module contains allophone categories rather than strictly phoneme ones, and if allophone categories with weak acoustic cues are recognized more slowly in the phoneme decision module.

This approach could become unwieldy if applied to all the allophonic variation of a language, though. There are a great many allophonic alternations, and also many smaller environmentally determined sound differences, and these would all have to be included in the lexical representation. Spinelli et al. (2003), who suggest a lexical representation mechanism, find allophonic processing effects based on differences in duration of /R/ depending on word boundary location, for example. Thus, a lexical representation mechanism that included all allophonic differences, even very low-level ones, would have to include very detailed lexical representations for every sound in every word, and an extreme number of “allophones” in the phoneme decision module. It would still differ from an exemplar model in that individual exemplars heard in the past would not be stored.

How an exemplar model might account for the non-word phoneme-monitoring data would depend on the types of units the model includes. If only whole words were recognized by the model, the appropriateness effect in non-words (1 above) could only be explained by reference to similar real words (e.g. [hok(i)to] partially activating exemplars of [k(i)ta] ‘North’), as in the Merge explanation above. This would fail to explain the limitation of the acoustic strength effect to non-word experiments. It might be possible to account for the results in an exemplar model with both allophones and words as categories, though. When recognizing

allophone-sized categories (e.g. [(i)]), the preference for appropriate allophones would reflect the preponderance of past exemplars with that allophone in similar environments (e.g. more exemplars of [(i)] in [k\_t] environment than in [dʒ\_g] environment). The overall preference for allophones with stronger acoustic cues could be modeled as higher baseline activation. To limit the acoustic strength effect to the phoneme-monitoring results, one would have to assume that word-level categories have no such higher activation for the form containing allophones with stronger acoustic cues. That is, the higher baseline activation for stronger allophones would only apply to allophone-level categories, which is a rather arbitrary stipulation. Exemplar models allow considerable flexibility in what the categories might be (Goldinger & Azuma, 2003; Johnson, 1997; Pierrehumbert, 2002), so such an adaptation of the model might be possible.

Thus, only two of the four models (lexical representation and exemplars) could account for the current results, even with modification. These two require an assumption that allophones with stronger acoustic cues or greater durations are processed more easily overall, and they require a way to limit this acoustic strength effect to sound-level or non-word processing. The Merge model already includes a separate sound-level decision module, making this limitation less arbitrary in that model. The similarity, across models, of the modifications needed leads to a further conclusion: it is not necessary for listeners to remember individual past exemplars in order to account for what are sometimes called “fine phonetic detail” effects. The lexical specification method assumes fine acoustic specification in the lexicon, but does not refer to individual past tokens. The current results are clearly a case of fine phonetic detail affecting processing, but memory for individual past exemplars is not necessary to model these results.

The past literature suggests that many factors may influence how allophones are processed (e.g. across word boundary vs. within word, neutralizing a distinction vs. not,

complete vs. partial assimilation (Gow & Im, 2004), progressive vs. regressive effects (Weber, 2001)). It is therefore quite possible that differing types of allophonic alternations involve differing mechanisms for processing. Japanese vowel reduction within a word, which does not neutralize any distinction (does not turn one phoneme into what sounds like another), presents a very different problem to the listener than across-word-boundary place assimilation does, so the listeners may reasonably use a different mechanism to process it. Recent evidence by Cutler et al. (submitted) indicates strongly that the Japanese reduced vowel alternation is processed at a lexical rather than a prelexical level, while Mitterer and McQueen (submitted) find evidence for prelexical processing of word-final consonant-cluster reduction. Specifically, Cutler et al. (submitted) find that listeners are more likely to assume a consonant-consonant string contains a reduced (deleted) vowel if the lexicon contains a greater number of words with vowel reduction between those two consonants. For example, listeners more easily assume a reduced /u/ is present in a string [ksa], for which there are relatively many words in the lexicon containing the string /kusa/ with reduction, than they do in a string /ʃha/, for which there are few words in the lexicon containing /ʃuha/ with reduction. Cutler et al. (submitted) argue that this, together with further evidence that Japanese listeners do not restore reduced vowels automatically in devoicing environments, shows that listeners restore Japanese reduced vowels through reference to lexical representations that specify vowel reduction. In the current data, the difference between lexical decision processing of real words and phoneme monitoring processing of non-words also leads to this conclusion. Since the results rule out the phonological inference model and the auditory processing model, this only leaves support for a lexical solution (either lexical listing of allophonic forms, or an exemplar-based lexical solution). Together, the current results and others' work indicate that Japanese vowel reduction is processed at the lexical level through lexical

representations that contain allophonic information, while some other types of allophonic variability are processed through other mechanisms. Listeners do not use a single mechanism for processing of all types of allophonic variability.

### **Conclusion**

The current study investigates listeners' processing of a particular type of allophonic variation, high vowel reduction in Japanese, through phoneme monitoring and lexical decision experiments. The results indicate that the strength of acoustic cues combines with knowledge of allophonic variation to affect recognition of sounds and words. The results show separate influences of listeners' language-specific knowledge of allophonic and phonotactic patterns. This study shows that in some cases, processing is actually facilitated by allophonic variation, even if the appropriate allophone is the one less similar to the underlying form and the one with weaker acoustic cues. The results can be modelled by putting detailed allophonic information into lexical listings, either through relatively abstract allophonic lexical representations or through an exemplar model, but they rule out an auditory processing or phonological inference model for processing of this alternation. It is likely that listeners employ more than one processing method because of the wide range of types of allophonic variation in human language.

## REFERENCES

- Cutler, A., Ooijen, B. van, Norris, D., & Sanchez-Casas, R. (1996). Speeded detection of vowels: A cross-linguistic study. *Perception and Psychophysics*, *58*, 807-822.
- Cutler, A., Otake, T., & McQueen, J. M. (Submitted). Vowel devoicing and the perception of spoken Japanese words.
- Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1568-1578.
- Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*, *16*, 491-505.
- Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science*, *19*, 407-439.
- Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 380-396.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251-279.
- Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305-320.
- Gow, D. W. Jr. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133-159.
- Gow, D. W. Jr. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 163-179.

- Gow, D. W. Jr., & Im, A. M. (2004). A cross-linguistic examination of assimilation context effects. *Journal of Memory and Language*, 51, 279-296.
- Hammond, M. (1999). *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. New York: Oxford University Press.
- Han, M. (1994). Acoustic manifestations of mora-timing in Japanese. *Journal of the Acoustical Society of America*, 96, 73-82.
- Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp.145-165). London: Academic Press.
- Kitahara, M. (1998). The interaction of pitch accent and vowel devoicing in Tokyo Japanese. In D. J. Silva, (Ed.), *Japanese/Korean Linguistics, Volume 8* (pp.303-315). Chicago: The University of Chicago Press.
- McQueen, J. M. (2007). Eight questions about spoken-word recognition. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics*. Oxford: Oxford University Press.
- Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception and Psychophysics*, 65, 956-969.
- Mitterer, H., Csépe, V., & Blomert, L. (2006). The role of perceptual integration in the perception of assimilated word forms. *Quarterly Journal of Experimental Psychology*, 59, 1395-1424.
- Mitterer, H., & McQueen, J. M. (Submitted). Processing reduced word forms in speech perception using probabilistic knowledge about speech production.
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299-370.

- Ostreicher, H. J., & Sharf, D. J. (1976). Effects of coarticulation on the identification of deleted consonant and vowel sounds. *Journal of Phonetics*, 4, 285-301.
- Patterson, D., LoCasto, P. C., & Connine, C. M. (2003). Corpora analyses of frequency of schwa deletion in conversational American English. *Phonetica*, 60, 45-69.
- Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp.101-139). New York: Mouton de Gruyter.
- Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, 57, 273-298.
- Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233-254.
- Vance, T. J. (1987). *An Introduction to Japanese Phonology*. Albany: State University of New York Press.
- Vance, T. J. (In press). *The sounds of Japanese*. Cambridge University Press.
- Weber, A. C. (2001). Help or hindrance: How violation of different assimilation rules affects spoken-language processing. *Language and Speech*, 44, 95-118.
- Yuen, C. L.-K. (2000). The perception of Japanese devoiced vowels. In A. Okrent & J. P. Boyle (Eds.), *Proceedings of the Chicago Linguistic Society*, v. 36-1 (pp. 531-547). Chicago: Chicago Linguistic Society.

### Figure Captions

Figure 1. Waveforms and spectrograms of a non-word /hokito/ with reduction of the vowel [(i)] (left) and with an unreduced vowel [i] (right). The reduced vowel has no periodic wave or formants, thus the vowel was deleted. Frication noise follows the burst of the [k].

Figure 2. A: Experiment 1 RT results (Japanese listeners, phoneme monitoring). Stars indicate statistically significant differences between reduced and unreduced vowels, and error bars display 95% confidence intervals. B: Percentage errors.

Figure 3. A: Experiment 2 RT results (American listeners, phoneme monitoring). Error bars are omitted because RTs are not analyzed statistically, as explained in the text. B: Percentage errors.

Figure 4. A: Experiment 3: RTs (ms) by Japanese listeners, with broader variety of voicing environments. B: Percentage errors.

Figure 5. Voicing environment RTs divided by preceding vs. following voiced consonant.

Figure 6. A: Experiment 4: RTs (ms) for Japanese listeners, for lexical decision. B: Percentage errors.

Figure 7. A: Experiment 5: RTs (ms) for Japanese listeners, for lexical decision with a broader range of voicing environments. B: Percentage errors.

Figure 8. Voicing environment RTs divided by preceding vs. following voiced consonant.

## Endnotes

---

<sup>1</sup> We will use the “(V)” notation for devoiced/reduced vowels throughout.

<sup>2</sup> Thus we refer to the environment as a “devoicing” or “voicing” environment, but the vowel itself as “reduced” or “unreduced” in order to avoid confusion about whether we are referring to the environment or the vowel.

<sup>3</sup> The sequence [dʒi] in Japanese is considered to consist of /zi/ phonemically (Vance, 1987).

However, we will write “dʒ” even in phonemic transcriptions to avoid switching between “dʒ” and “z.”

## Appendix A: Stimuli and acoustic information for Experiments 1 and 2 (phoneme monitoring).

## Devoicing Environment

hokito	sekite	nuhika	seshita	moshito	meshite
wachika	machike	nokita	mahiko	tekisa	kachiho
nehisa	mukito	yuchite	sumokika	toyakiko	nagahita
kutekito	hasechito	tadahika*	saneshita	moneshito	wagashite*
yawachike	motahike	nosahiko	kotashike	notochiko	menahisa

## Voicing Environment

tajida	nujida	wajide	tejido	yajido	wajina
tajiga	tejiba	kojiba	sojime	mojiza	kejizo
hojido	mejina	yujima	natajiba	kunojime	sasajina
ketajiba	mokojibe	narajizo	merajime	warejida	masujido
kanajida	warejina	tasajiba	yatejima	yotajino	towajina

## Nasal Environment

wanide	tenido	saniza	kunima	nanime	tanina
tanimo	yunida	waniba	nenigo	soniba	hanina
sunigo	tonize	meniga	hasanina	yosanine	kedanida
nasaniba	fukonino	ketanime	kutonino	wasenida	tarenido
tasonibe	kasonide	samonina	yuranide	nuraniza	wadonima

Pitch accent for all items was realized with accent on the antepenultimate mora, except for starred items, for which the reduced version only was realized as unaccented.

Durations: Total stimulus duration and target mora duration are given for all categories, in milliseconds. For the unreduced stimuli only, the duration of the target [i] and of its preceding consonant can be measured separately.

Environment	Reduced	Stim. dur.	Target mora dur.	Unreduced	Stim. dur.	Target mora dur.	Preced. C dur.	[i] dur.
Devoic.	3-mora	494	132	3-mora	519	145	60	85
	4-mora	611	120	4-mora	678	145	66	79
Voicing	3-mora	471	110	3-mora	476	121	42	78
	4-mora	591	99	4-mora	630	113	49	64
Nasal	3-mora	469	158	3-mora	458	159	51	108
	4-mora	571	144	4-mora	578	140	51	90

In the nasal environment, for the 7 stimuli containing [ɲ(i)ɲ] the reduced version had one long nasal ([hɲɲa]) in the spectrogram. Identification of the target mora ([ɲ(i)]) separately from the following the alveolar nasal was achieved primarily by listening rather than by spectral cues.

## Appendix B: Materials and acoustic information for Experiment 3.

## Devoicing Environment

hokito	seshta	moshito	meshite	wachika	machike
mahiko	tekisa	kachiho	mukito	nagahita	hasechito
tadahika	moneshito	wagashite	yawachike	motahike	kotashike
notochiko	menahisa				

## Voicing Environment

wajide	teshido	wajina	tajiga	teshiba	toshiba
mojiza	kekizo	hojido	mechina	kanajida	kunoshime
sasajina	ketajiba	mokojibe	natakiba	merahime	wareshida
tanachida	tasajiba				

## Nasal Environment

tenido	saniza	kunima	nanime	tanina	tanimo
soniba	sunigo	tonize	meniga	kedanida	nasaniba
fukonino	ketanime	wasenida	tarenido	tasonibe	kasonide
samonina	yuranide				

Pitch accent patterns for the devoicing environment were H(L)L for 3-mora stimuli and LH(H)H or LH(L)L for 4-mora stimuli. For the voicing and nasal environments, some 3-mora stimuli had were L(H)H. The pitch patterns were identical between the reduced vowel stimuli and the full vowel stimuli except the lack of pitch for the reduced vowel.

Durations: Total stimulus duration and target mora duration after splicing are given for all categories, in milliseconds. For the unreduced stimuli only, the duration of the target [i] and of its preceding consonant can be measured separately.

Environment	Reduced	Stim. dur.	Target mora dur.	Unreduced	Stim. dur.	Target mora dur.	Preced. C dur.	[i] dur.
Devoic.	3-mora	521	140	3-mora	536	156	60	95
	4-mora	626	128	4-mora	658	160	62	98
Voicing	3-mora	496	145	3-mora	525	167	60	108
	4-mora	599	129	4-mora	609	137	53	84
Nasal	3-mora	473	160	3-mora	484	170	52	117
	4-mora	585	151	4-mora	602	168	52	117

## Appendix C: Materials and acoustic information for Experiment 4.

## Devoicing Environment

<i>akikan</i> ‘empty can’	<i>hashika</i> ‘measles’	<i>kachiku</i> ‘domestic animals’
<i>tsukiasu</i> ‘stub’	<i>akisu</i> ‘robber’	<i>oshikakeru</i> ‘go uninvited’
<i>ashita</i> ‘tomorrow’	<i>fukitsu</i> ‘ill omen’	<i>yakitori</i> ‘grilled chicken’
<i>onshitsu</i> ‘a green house’	<i>koshitsu</i> ‘a single room’	<i>hakike</i> ‘nausea’
<i>mushisasare</i> ‘insect bite’	<i>kakikomu</i> ‘to write down’	<i>sekikomu</i> ‘to cough’
<i>ochikomu</i> ‘to get depressed’	<i>yakisoba</i> ‘fried noodle’	<i>soshitsu</i> ‘nature’
<i>tekitoo</i> ‘reasonable’	<i>washitsu</i> ‘Japanese-style room’	

## Voicing Environment

<i>mijikai</i> ‘short’	<i>ojisan</i> ‘uncle’	<i>kujibiki</i> ‘lottery’
<i>akijikan</i> ‘free time’	<i>kejime</i> ‘to be distinguishable’	<i>nodojiman*</i> ‘singing contest’
<i>genjitsu</i> ‘real world’	<i>kujikeru</i> ‘to be discouraged’	<i>hajimeru</i> ‘to begin’
<i>tejina</i> ‘conjuring tricks’	<i>majime</i> ‘serious’	<i>fujisan</i> ‘Mt. Fuji’
<i>tatejima</i> ‘vertical stripe’	<i>hajiku*</i> ‘to snap’	<i>najimu*</i> ‘to become familiar’
<i>mijime</i> ‘miserable’	<i>kokugojiten*</i> ‘Japanese dictionary’	
<i>nekojita</i> ‘sensitive to hot food’		

## Nasal Environment

<i>onigiri*</i> ‘rice ball’	<i>tanigoe</i> ‘over a valley’	<i>tenimotsu</i> ‘luggage’
<i>gyuuniku</i> ‘beef’	<i>nanimono</i> ‘whoever’	<i>wanigawa</i> ‘crocodile skin’
<i>ganimata</i> ‘bowlegged’	<i>inisharu*</i> ‘initial’	<i>tanima</i> ‘ravine’
<i>kaniza</i> ‘Cancer sign’	<i>onigokko</i> ‘playing tag’	<i>butaniku</i> ‘pork’
<i>kunizukuri</i> ‘to form a nation’	<i>aniki</i> ‘elder brother’	<i>banira</i> ‘vanilla’
<i>hiniku</i> ‘sarcasm’	<i>minikui</i> ‘ugly’	<i>monitaa</i> ‘monitor’
<i>yonige</i> ‘to flee by night’	<i>kuinige</i> ‘to run away without paying one’s bill’	

Pitch patterns of all stimuli were appropriate for the Tokyo dialect according to the NHK Accent Dictionary (1985). Pitch patterns for reduced and unreduced stimuli were identical except for the lack of pitch in reduced vowels. Starred stimuli had a reduced vowel in an accented syllable.

Durations: Total stimulus duration and target mora duration are given for all categories, in milliseconds. For the unreduced stimuli only, the duration of the target [i] and of its preceding consonant can be measured separately.

Environ- ment		Stim. dur.	Target mora dur.		Stim. dur.	Target mora dur.	Preced. C dur.	[i] dur.
Devoic.	Reduced	616	130	Unreduced	654	165	66	99
Voicing	Reduced	581	122	Unreduced	609	137	46	91
Nasal	Reduced	553	190	Unreduced	626	187	68	119

## Appendix D: Materials and acoustic information for Experiment 5.

## Devoicing Environment

<i>akikan</i> ‘empty can’	<i>hashika</i> ‘measles’	<i>kachiku</i> ‘domestic animals’
<i>akisu</i> ‘robber’	<i>oshikakeru</i> ‘go uninvited’	<i>ashita</i> ‘tomorrow’
<i>yakitori</i> ‘grilled chicken’	<i>onshitsu</i> ‘a green house’	<i>koshitsu</i> ‘a single room’
<i>hakike</i> ‘nausea’	<i>ochikomu</i> ‘to get depressed’	<i>yakisoba</i> ‘fried noodle’
<i>soshitsu</i> ‘nature’	<i>tekitoo</i> ‘reasonable’	<i>kakikotoba</i> ‘written language’
<i>sekitori</i> ‘sumo wrestler’	<i>ekitai</i> ‘liquid’	<i>sekihan</i> ‘red rice’
<i>machikado</i> ‘street corner’	<i>washitsu</i> ‘Japanese-style room’	

## Voicing Environment

<i>mijikai</i> ‘short’	<i>ojisan</i> ‘uncle’	<i>takibi</i> ‘bonfire’
<i>soojiki</i> ‘vacuum cleaner’	<i>kejime</i> ‘to be distinguishable’	<i>tachiba</i> ‘standpoint’
<i>ichigo</i> ‘strawberry’	<i>kujikeru</i> ‘to be discouraged’	<i>akirameru</i> ‘to give up’
<i>tejina</i> ‘conjuring tricks’	<i>majime</i> ‘serious’	<i>sashidasu</i> ‘stretch out’
<i>hajiku*</i> ‘to snap’	<i>techigai*</i> ‘mistake’	<i>sekidome</i> ‘cough syrup’
<i>sashimi</i> ‘sliced raw fish’	<i>oshidasu</i> ‘to push out’	<i>dekigoto*</i> ‘incident’
<i>daijiken*</i> ‘big incident’	<i>nekojita</i> ‘sensitive to hot food’	

## Nasal Environment

<i>onigiri*</i> ‘rice ball’	<i>tanigoe</i> ‘over a valley’	<i>tenimotsu</i> ‘luggage’
<i>gyuuniku</i> ‘beef’	<i>tenisu</i> ‘tennis’	<i>wanigawa</i> ‘crocodile skin’
<i>haniwa</i> ‘clay figure’	<i>inisharu*</i> ‘initial’	<i>tanima</i> ‘ravine’
<i>kaniza</i> ‘the Cancer’	<i>onigokko</i> ‘playing tag’	<i>butaniku</i> ‘pork’
<i>kunizukuri</i> ‘to form a nation’	<i>aniki</i> ‘elder brother’	<i>banira</i> ‘vanilla’
<i>hiniku</i> ‘sarcasm’	<i>minikui</i> ‘ugly’	<i>monitaa</i> ‘monitor’
<i>yonige</i> ‘to flee by night’	<i>kuinige</i> ‘to run away without paying one’s bill’	

Pitch patterns of all stimuli were appropriate for the Tokyo dialect according to the NHK Accent Dictionary (1985). Pitch patterns for reduced and unreduced stimuli were identical except for the lack of pitch in reduced vowels. Starred stimuli had a reduced vowel in an accented syllable.

Durations: Total stimulus duration and target mora duration after splicing are given for all categories, in milliseconds. For the unreduced stimuli only, the duration of the target [i] and of its preceding consonant can be measured separately.

Environment	Stimuli	Stim. dur.	Target mora dur.	Stimuli	Stim. dur.	Target mora dur.	Preced. C dur.	[i] dur.
Devoic.	Reduced	615	136	Unreduced	645	167	60	107
Voicing	Reduced	603	145	Unreduced	618	159	54	94
Nasal	Reduced	570	188	Unreduced	582	200	67	119