Japanese Listeners’ Perceptions of Phonotactic Violations

Laurel Fais¹, Sachiyô Kajikawa²,³, Janet Werker¹ and Shigeaki Amano²

¹ University of British Columbia, Vancouver
² NTT Communication Science Laboratories, Keihanna, Japan
³ Tamagawa University Research Institute, Tokyo

Abstract

The canonical form for Japanese words is (Consonant)Vowel(Consonant)Vowel⁻. However, a regular process of high vowel devoicing between voiceless consonants and word-finally after voiceless consonants results in consonant clusters and word-final consonants, apparent violations of that phonotactic pattern. We investigated Japanese adults’ perceptions of these violations, asking them to rate both canonical and noncanonical nonsense forms on a scale of goodness. Results indicate that adults show evidence of being guided in making their judgments by an implicit understanding of both typical canonical forms and appropriate contexts for vowel devoicing.

1 Introduction

How listeners perceive speech that is not in their native language is a question that has received considerable attention (Best, 1994; Kuhl, 1993; Werker, 1995; Werker & Tees, 1999). But what of the perception by native speakers of acoustic forms in their own language that violate basic phonotactic constraints and are thus, in some sense, “non-native”? In this work, we examine the case of acoustic forms in a language in which regular phonological processes obscure canonical phonotactic patterns. That is, processes operating “lawfully” produce acoustic forms that violate general phonotactic constraints. Will native listeners perceive these types of words according to the explicit acoustic form that results from the operation of these processes, or according to their implicit canonical form?

We investigated this question by looking at how Japanese listeners perceive nonsense word forms with and without apparent final consonant clusters.

Acknowledgments: We gratefully acknowledge the help of Nenagh Kemp and Laura Sabourin with the finer points of our statistical analysis. This work was partially funded by grants to the third author from the Human Frontier Science Program and the Social Sciences and Humanities Research Council of Canada.

Address for correspondence. Laurel Fais, Department of Psychology, UBC, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada; e-mail: <lfais@psych.ubc.ca>.
1.1 Japanese phonotactics and vowel devoicing

The predominant canonical form for morae in Japanese is (Consonant)Vowel [(C)V], so that words in Japanese tend to be CVCV− (e.g., *tabemono* food). In particular, the only allowable word-internal “consonant clusters” actually involve a moraic nasal followed by a consonant (e.g., *manga* comic book). We did not use moraic nasals in our stimuli, and so will assume that (C)V− is the canonical form for the work reported here. Because all nonmoraic consonants must be followed by a vowel, there are no word-final consonant clusters.

There is also a phonological process of vowel devoicing that operates in Japanese. Briefly, high vowels (/ɯ/ and /i/) may be devoiced if they occur between two voiceless consonants or after a voiceless consonant word finally. The process of vowel devoicing may result in a shortened, partially voiced vowel or in the complete devoicing of the vowel. In the latter case, if the vowel follows a plosive, the “vowel” may be realized as a longer period of aspiration after the consonant or as the release of the plosive into a homorganic fricative. If the “vowel” follows a fricative, it is realized as a brief extension of the frication for the consonant (Kondo, 1994; Shibatani, 1990). The resulting percept is that of two consecutive consonants, which may then be interpreted by non-native speakers as a consonant cluster (Kondo, to appear). For example, /kiɿa/ (train) is commonly pronounced in conversational speech as /kɿa/, but understood as /kiɿa/.

High vowels almost always devoice in a devoicing context, unless two or more such contexts occur consecutively. In that case, devoicing in both contexts is optional, and devoicing of three vowels in consecutive environments is highly unlikely (Kondo, to appear). Rates of devoicing are higher for unaccented vowels, vowels in word-initial syllables, in contexts that do not fall at a morpheme boundary, and at fast reading rates (Kuriyagawa & Sawashima, 1989; Sugito, to appear; Yoshida, to appear). Thus, vowel devoicing rates are variable, and voiced and devoiced versions of vowels occur with differing frequencies depending upon context, from virtually zero frequency for low vowels between voiced consonants to near 100% devoicing of high vowels between voiceless consonants in word-initial syllables in rapid speech (we examine devoicing contexts and frequencies in more detail in the Discussion section below). In this work, we are investigating whether listeners’ perceptions are guided primarily by their understanding¹ of the phonotactic regularities of Japanese or by their understanding of these differing frequencies for devoiced vowels in differing contexts.

---

¹ Where we use ‘understanding’ in this paper, no conscious thought process is implied. We simply prefer to use the term ‘understanding’ instead of ‘knowledge’, which may convey the impression of the existence ‘in a hearer’s head’ of some type of static representation. Our data do not allow us to make any claims regarding the form of knowledge about language “in a hearer’s head.”
1.2 Perception of vowel devoicing

Morris (2004) asked native adult speakers of Japanese to listen to Japanese speakers pronounce words in a list and judge whether those speakers were from the Kanto or Kinki areas. Kanto area speakers are said to have a higher rate of devoicing than Kinki area speakers (Imaizumi, Fuwa, & Hosoi, 1999). Some of the words were four-mora words with one vowel in a devoicing context (the rest of the words were prepared to test listeners' perceptions of pitch accent and do not concern us here). Two tokens were prepared for each of these words: one with the vowel voiced and one with the vowel devoiced. She found that listeners judged speakers saying these words to be Kanto speakers, whether or not the vowel was actually devoiced. Morris’s results seem to indicate that listeners do have a sense of the contexts in which vowels may be frequently devoiced, based upon their experience with such contexts in their speech input.

Dupoux and colleagues investigated Japanese adult listeners’ perception of artificially created consonant clusters in word-internal position in nonsense words (Dupoux, Fushimi, Kakehi, & Mehler, 1999; Dupoux, Kakehi, Hirose, & Pallier, 1999). Crucially, the position from which the vowel was excised to create the clusters was between two voiced consonants, a position in which vowels are virtually never devoiced. Their results indicated that a majority of adult listeners of Japanese perceived an epenthetic high vowel /ɯ/ between voiced consonants word-externally where, in fact, all vocalic content had been excised. Their interpretation of this finding is that these listeners’ perceptions were shaped by the phonotactic constraints of Japanese such that the resulting percept followed the canonical (C)V(C)V~ form. But, in fact, Japanese listeners could just as easily have been basing their judgments on the fact that, since this is an extremely infrequent context for vowel devoicing, they have no experience with this acoustic form in their native language input.

Morris’s (2004) result, namely, that listeners judged that a speaker spoke a devoicing dialect even when the vowel was actually voiced, is the converse of the result of Dupoux and colleagues: in the latter case, listeners reported hearing epenthetic vowels where none actually occurred; in the former case, listeners judged speech on the basis of a lack of vowels that were sometimes actually present. Morris’s results reveal a sensitivity of listeners to possible acoustic forms rather than to prototypical canonical forms. Dupoux and colleagues interpret their results as indicative of an understanding of phonotactic constraints.

1.3 Acoustic forms or phonotactic constraints?

In this work, we extend the investigation of the mechanisms that guide listeners’ perceptions by focusing on a word-final, post-voiceless-consonant context, using Japanese nonsense words. This is a position in which high vowels may be devoiced,

2 However, see (Maekawa & Kikuchi, to appear) for a rare corpus-based analysis of vowel devoicing in which they report devoicing even for nonhigh vowels between voiced consonants, albeit at very low rates (/a/: 0.15%; /ɛ/: 0.25%; /ɔ/: 0.08%).
resulting in a final mora of the form ~C# (where ‘#’ denotes the end of a word), an apparent violation of the canonical ~CV# form. We used a standard goodness rating task to probe how an understanding of phonotactic constraints and an understanding of the frequency of occurring acoustic forms interact in guiding the perception of vowels in this context. Listeners were presented with Japanese nonsense words ending in contexts in which vowel devoicing was permitted or not permitted, and in which vowels were present or absent. If phonotactics alone guide perception, as Dupoux and colleagues suggest, we expected that listeners would perceive word-final vowels in both word-final contexts, regardless of whether the vowel was actually present or not. This would mean that they would rate canonical and noncanonical forms similarly. This reasoning is somewhat counterintuitive. It would seem more straightforward to claim that, if listeners judge forms on the basis of phonotactics, they would prefer canonical forms and disprefer noncanonical forms, rather than rating them both similarly. The conclusion that they will rate them similarly is the natural consequence of the proposal that hearers “create” canonical forms from noncanonical input (à la Dupoux, Fushimi et al., 1999). Note that this “creation” is quite explicitly attested in Dupoux’s study, since listeners wrote down what they thought they heard, and included in their reports overt vowels that in fact were not present in the input. Thus, if listeners are forming their judgments strictly on the basis of their implicit understanding of phonotactics, they will rate the canonical and noncanonical forms equally. On the other hand, if frequency of acoustic forms plays a primary role, listeners should only “hear” a vowel when it is actually present in the signal, regardless of the context. In this interpretation, then, listeners will rate forms that they are more likely to have actually heard, as more preferred forms.

2 Methods

Participants were 40 Japanese native speakers: 20 speakers who had grown up in the Kanto area (roughly, Tokyo and environs: Kanagawa, Chiba, Saitama, Ibaraki, Tochigi, Gunma, Yamanashi); 10 males and 10 females; mean age 26.5, range = 21 – 29, $SD = 2.85$; and 20 speakers who had grown up in the Kansai area (roughly, Kyoto and environs: Osaka, Hyogo, Mie, Nara, Wakayama); 10 males and 10 females; mean age 24.8, range = 21 – 29, $SD = 2.33$. None of the participants had ever lived abroad nor had they lived with native English speakers. However, one of the Kanto participants had lived in Osaka from 0 to 1 year old, and another had lived in Hiroshima from 4 to 6 years old. Two of the Kansai participants had lived in the Kanto area for similar periods. Their responses are included in the results below.

The stimuli were eight nonsense words: neek [ni:k], nee ku [ni:kui], neeks [ni:ks], neekusu [ni:kusɯ], kee t [ki:t], keeto ’ki[to], keets [ki:ts], and keetsu [ki:tsu] (Table 1). The first mora of each word was accented; this is the most natural accent pattern for nonsense words of this form. Five intonational variations of each nonsense item were recorded by a completely fluent, female, bilingual, English-Japanese speaker of the Kinki dialect. The F0 patterns of those variations were matched across the eight items, and were typical of interested, casual speech (i.e., they were not in citation or list intonation). It was necessary to use a bilingual speaker so that the pronunciations of the Japanese forms were as native as possible, while the pronunciations of
the word-final consonants did not contain residual vowel production, difficult for a Japanese speaker to accomplish in careful speech because of the canonical \~CV# form of Japanese words. Informal rating of the tokens by three native speakers of Japanese (1 a speaker of the Kanto dialect and 2 Kinki dialect speakers) determined the canonical forms to be perfectly “Japanese-like.”

The words were chosen to allow us to examine a number of different factors. The pairs neek/neeku and keets/keetsu exemplify the word-final context in which the vowel may be devoiced, after a voiceless stop in the former pair and a voiceless affricate in the latter pair (it is important to note that [ts] is an affricate and thus a single consonant in Japanese). The pair consisting of keet/keeto is an example of a word-final context in which the vowel may not be devoiced, since /o/ is not a high vowel. While keets is a possible acoustic form, that is, derived from keetsu by final vowel devoicing after a voiceless C, keet is not. Before high vowels, /t/ has the allophonic form of either [tɿ] (only before /i/) or [ts] (only before /u/). Thus, an acoustic form like keet could not have been derived by vowel devoicing, a process that affects only high vowels. We included the pair neeks/neekusu as well, to investigate the effect of multiple applications of the devoicing process in consecutive devoicing contexts. The acoustic form neeks could be derived as a result of the concurrent vowel devoicing of the /u/ occurring between the voiceless consonants /k/ and /s/ and of the word-final /u/ occurring after the voiceless fricative.

Table 1
Stimuli

<table>
<thead>
<tr>
<th>Canonical</th>
<th>Noncanonical</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVCV(CV)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| No devoicing context | ki:to |       |
| Devoicing context    | ki:tsu | ki:ts |
|                       | ni:ku | ni:k  |
|                       | ni:ku:SU | ni:ks |

The apparatus used was a laptop PC (Windows), an amplifier (STAX SRM), and headphones (STAX SR_PRO).
Each group of two or three participants rated the utterances in a sound proof chamber. Sound stimuli were presented through the PC, amplifier and headphones to each participant individually. The orthographic form of the item against which the auditory token was to be evaluated appeared on the computer monitor (Fig. 1). The items were written in *katakana*, the standard syllabary for writing foreign or, in this case, nonsense words. The instructions that each participant received prior to the start of the experiment stipulated that they should (1) evaluate how good an exemplar of the displayed word the sound presented was; (2) ignore syllable stress or intonation in their evaluations; (3) try to use all the points on the scale throughout the experiment; (4) determine standards of evaluation for themselves for each of the seven points during the initial practice sessions.

**Figure 1**
The screen seen by participants in the study

After being instructed about the procedure, the participant clicked the start button on the display. A nonsense word written in *katakana* and a rating scale were shown on the display and a sound stimulus was played once. The auditory tokens of all items were evaluated against the written forms for the nearest canonical form. Thus, the auditory tokens [ki] and [kitto] were evaluated against the orthographic representation for (the canonical Japanese form) *keeto*; [kits] and [kitsu] were evaluated against the orthographic form for (the canonical) *keetsu*; [ni:k] and [ni:ku] against *neeku*; and [ni:ks] and [ni:kusu] against *neekusu*.

The rating scale consisted of seven points, where Point 1 designated “a bad example of the Japanese word shown on the display” and Point 7 designated “a good example of the Japanese word shown on the display.” The participant clicked on a score, and then clicked the “next” button to listen to and rate the next sound.

There were four sessions for each participant. The first three sessions were practice sessions in which 10 tokens of the stimuli were rated. A total of 30 tokens (10
tokens × 3 sessions) were used randomly for the practice sessions. In the fourth session, the participants rated 200 tokens (8 items × 5 intonational variations × 5 times).

3 Results

Figure 2 shows the mean ratings of all the word types by speakers of Kanto and Kinki dialects. We conducted a 2 (dialects: Kanto and Kinki) × 8 (word types) mixed ANOVA. The main effect of dialect was not significant, $F(1, 38) = 1.38, p = .16$; however, there was a significant main effect of word type in the ratings, $F(7, 266) = 187.85, p < .001$. In post hoc tests, Fisher’s LSD was used to examine the differences among word types; those results are given below (Figures 3–6). Note that the ratings were significantly higher for each of the canonical forms when compared to the corresponding noncanonical form ($p < .01$). The interaction of word and dialect was not significant, $F(7, 266) = 0.52, p = .47$.

Figure 2
Overall results for mean goodness ratings of all eight nonsense word forms, for each dialect (bars denote $SD$)

It is interesting to look in more detail at some of the natural subsets of this data. One such group consists of keet, keets, and neek, all noncanonical forms that lack a final voiced vowel. Recall that keet is not a possible acoustic form, while keets and neek are possible forms resulting from devoicing of the final vowel (Table 1). Figure 3 shows the mean ratings for these items.

The rating for neek, a possible acoustic form, was significantly higher than that for keet, an impossible one ($p < .004$). The rating for keets was significantly higher than that for keet or neek ($p < .001$).
Conversely, we can also examine the voiced vowel equivalents of the items in Figure 3, that is, forms in which voiced vowels appear in a context in which they may be legally devoiced. In *keeto*, the final vowel can never be devoiced, while in *keetsu* and *neeku*, final vowels appear in a potentially devoiced context. Figure 4 gives the mean ratings for these forms.

The differences among these three ratings are all significant: *neeku* is rated more highly than *keeto* ($p < .002$) and *keetsu* ($p < .001$), and *keeto* is rated more highly than *keetsu* ($p < .001$).

The inclusion of the form *neeks* in the stimulus set allows us to compare forms in which there is only one context for devoicing, that is, *keets* and *neek*, to a form in which there are two adjacent contexts for devoicing, that is, *neeks*. Figure 5 gives these ratings.

---

3 There are two possible contributing factors to the somewhat low ratings for the canonical forms. The fact that the stimuli were spoken with varied intonation and not in citation form might have made them sound less “usual.” In addition, it has been demonstrated that Japanese respondents to Likert scales such as this are more “centrist,” that is, tend to use more neutral ratings than European-North American respondents (Chen, Lee & Stevenson, 1995).
The rating for *keets* is significantly higher than that for *neeks* \( (p < .001) \). However, the opposite is true for *neek*; *neek* is rated significantly lower than *neeks* \( (p < .001) \).

We can look at the converse of this case as well, by comparing forms in which there is a voiced vowel in only one devoicing context, that is, *neeku* and *keetsu*, to a form in which there are voiced vowels in two potential devoicing contexts, that is, *neekusu*. Figure 6 gives the results for this comparison.

*Neekusu* is rated more highly than *keetsu* \( (p < .008) \). However, *neekusu* was rated significantly lower than *neeku* \( (p < .001) \). In fact, *keetsu* was also rated significantly lower than *neeku* \( (p < .001) \).

### 4 Discussion

Our results give us insight not only into listeners’ perceptions of the various word forms, but also into the phenomena of dialect and consecutive devoicing, as well as into the question of the interaction between phonotactics and perception of surface forms. We will take up each of these areas in turn.
4.1
A Comparison of Neeku and Keetsu

Let us first examine the results for the forms neeku and keetsu. We might assume that since both these forms contain a word-final vowel devoicing context, they would pattern similarly in the data. By similar reasoning, neek and keets should pattern similarly. However, in fact, the pairs neeku/keetsu and neek/keets have significantly different ratings. Further, the difference in the ratings for neek and keets is greater than the difference in the ratings for neek and keet, a form without a final vowel devoicing context at all (Fig. 3). Finally, as noted, neeku is significantly better than keetsu, but, unlike keetsu, it is also better than keeto (Fig. 4).

These data may be explained by examining the phonetic contexts of devoicing more closely. There are claims in the literature that a (voiceless) fric ative preceding a high vowel and a (voiceless) stop following the high vowel provide the context most conducive to devoicing, with /i/ and /u/ being devoiced in that context 98.4% and 97.5% of the time, respectively (Maekawa & Kikuchi, to appear; see also Yoshida, to appear). Other combinations of the manners of the preceding and following consonants are progressively less predictive of devoicing, with the lowest rate found between two voiceless fricatives; a devoicing rate of 27.3% between /s/ and /s/ is reported by Yoshida (to appear); and a rate of 61.2% for /u/ between voiceless fricatives and 38.1% for /i/ between voiceless fricatives is reported by Maekawa and Kikuchi (to appear). Although there is not exact agreement among studies that have investigated the influence of manner of articulation on the voicing of a vowel in a devoicing context, and although none of these studies has focused on devoicing of word-final vowels, we can take their overall results to be suggestive at least. There does seem to be agreement that a preceding fricative promotes, while a preceding stop discourages devoicing of the following vowel (Maekawa & Kikuchi, to appear; Yoshida, to appear). If we extrapolate this finding to the word-final context, it might be that the final frication of keets is a better promoter of devoicing than is the final stop, /k/ of neek. If we say, then, that neeku contains a context less conducive to final vowel devoicing than keetsu, the fact that these forms do not pattern together is no longer unexpected. Similarly, the fact that neek and keet, as well as neeku and keeto, seem to pattern similarly can be explained by the fact that neeku contains a context not particularly conducive to devoicing, while, likewise, keeto contains a context in which devoicing rarely if ever occurs. Like keet, then, we expect neek to be worse than keets because neek does not have a context particularly conducive to final vowel devoicing and keets does (Fig. 3); further, we expect neeku to be better than keetsu since keetsu could have an acoustic form in which its final vowel is devoiced (Fig. 4), and neeku is less likely to. These are, in fact, the results that we obtained.

There is one piece of data that is not accounted for by this understanding of neeku and keetsu, and that is the fact that neeku is rated more highly than keeto (Fig. 4). Because neeku contains a devoicing context in which a voiced vowel appears, it should be rated lower than keeto, which contains a context in which the final vowel is virtually never devoiced, and thus the voiced vowel is by far the most likely pronunciation. The fact that neeku is actually rated more highly than keeto remains unexplained. There is a possibility that frequency or distributional effects come into play in this...
situation (/kɯ/ may be heard more frequently than /to/, word-finally, for example), but these speculations are beyond the scope of this paper.

4.2 Dialect

It is interesting that dialect did not have a significant effect on the ratings. Speakers from the Kanto (Tokyo) area are claimed to have a higher rate of devoicing than speakers from the Kinki (Osaka-Kyoto) area (Imaizumi, Fuwa, & Hosoi, 1999). Nevertheless, speakers from these two areas rated the nonsense words in this study similarly. One possible explanation is that the two dialects do not, in actual fact, differ greatly in devoicing rates. In a survey of empirical results concerning devoicing rates in Kanto and Kinki dialects (Sugito, 1969, 1988; Tahara, Kishie, Gun, Tsuzome, Nakai, Maeda, & Muranaka, 1998; Yoshioka, 1981), Morris (2004) suggests that, although devoicing rates for the two dialects are culturally perceived to differ, they are, according to the acoustic analyses done in the studies she surveys, quite comparable. If this is indeed the case, these results are not so surprising.

But even if significant dialectal differences do exist in production of devoiced vowels, speakers of the two dialects in this study may have rated their perceptions of the word forms similarly as a result of cultural factors. The “standard” news commentator dialect in Japan is the Kanto dialect; certainly Kinki dialect speakers would be highly familiar with this dialect (and are even encouraged to learn to speak it in school at least, and often in the home as well; Shibatani, 1990). These speakers may have interpreted the forms lacking vowels in devoicing contexts as Kanto-like, and thus evaluated them as (they perceive) a Kanto speaker would, showing the sort of “reference-group effect” that plagues studies involving cross-cultural (here, cross-dialectal) subjective judgments (Heine, Lehman, Peng, & Greenholtz, 2002). This interpretation might have spread to affect the way the participants approached all of the tokens in the study. Indeed, given the fact that these stimuli were spoken as single words (i.e., not in a context of running speech), and were compared to a written representation, the participants may have been biased from the start to view them as “standard dialect,” that is, Kanto, forms. This initial bias would reduce the dialectal differences found in natural, spoken Japanese.

4.3 Phonotactics and Acoustic Forms

4.3.1 Canonical and noncanonical forms

Recall that we proposed two possible influences on the perception of the forms in this study, with differing predictions: If perceivers are judging the stimuli on the basis of the canonical forms for these items, they might perceive vowels to be present even when they are not (just as the listeners did in the voiced consonant context in the work by Dupoux and colleagues). In that case, their judgments for pairs with and without vowels present will be the same. The other possible influence is that of the perception of surface acoustics. If listeners are judging the stimuli on the basis of
what they actually hear, they will differentiate canonical forms from noncanonical forms and rate the former more highly.

The across-the-board significant difference in acceptability of canonical and noncanonical forms argues strongly against the interpretation that listeners are “supplying” an epenthetic vowel in these noncanonical contexts, as Dupoux and colleagues suggested for their participants in the context of voiced consonants. In the word-final context, listeners do not appear to perceive an epenthetic vowel that renders vowelless forms as phonotactically acceptable as their “voweled” counterparts. Rather, listeners appear to pay attention to acoustic forms and to judge those that do contain final consonant clusters to be much “worse” than those that do not.

Recall that we included an additional factor to be considered; the forms we used as stimuli are not simply canonical and noncanonical. Two of the noncanonical forms, keets and neek, are acceptable acoustic forms deriving from a process of devoicing, differing in acceptability based on the manner of the final consonant, and the third, keet, is not an acceptable form. A key question is: how were the noncanonical, but possible, acoustic forms keets and neek perceived? If an implicit understanding of phonotactics provides the primary criterion for judgment, the fact that some noncanonical forms do appear in the language input and others do not should have no effect; possible or not, all noncanonical forms will be perceived as canonical. As we have seen, this was not the pattern of results that we found.

On the other hand, if familiarity with acoustic forms provides the criterion for the perception of these word forms, we expect a different pattern of results. We predict that, if they are guided by acoustics, listeners will perceive forms based on the frequency with which they have heard those forms in the language around them. Certainly canonical forms will sound better to listeners than noncanonical forms because the former are the far more frequent, standard acoustic ones. But listeners’ judgments of both canonical and noncanonical forms will also be affected by listeners’ understanding of the varying possibilities for devoicing in particular contexts, an understanding that derives from a familiarity with the frequencies with which certain acoustic patterns occur in the language input. Guided by their knowledge of the likelihood of occurrence of possible acoustic forms, perceivers do note the differences between canonical and noncanonical forms, but they note differences among canonical forms and among noncanonical forms as well. We discuss these finer-grained distinctions next.

4.3.2 Single devoicing contexts

Listeners in this study judged keets to be better than neek, and neek to be better than keet (Fig. 3). This is a puzzling result if perceivers are judging these items based strictly on phonotactics. In that case, listeners should judge all three forms to be equally good, since they are all equally removed from the canonical form. Instead, we suggest that this result supports the notion that perceivers are guided by the acoustics of the forms, and are sensitive to frequencies of word-final consonants, that is, to the contexts of word-final devoicing. Thus, since [tsui]# is likely much more frequent in the casual productions of native Japanese speakers than [k]# and certainly more
frequent than [t]#, a form like keets is rated more highly than neek or keet. Further, listeners also judged neeku and keeto to be better than keetsu (Fig. 4). If phonotactics guided their perception, there would be no reason to judge one canonical form better than another. Rather, then, we suggest that perceivers are basing their judgments on the fact that, since /tsɯ/# has the two phonetic variants [tsɯ]# and [ts]#, listeners are more uncertain about which phonetic form to prefer than in the case of /kɯ/#, which does not have a very frequent corresponding phonetic variant [k]#, and /to/#, which does not have a corresponding phonetic variant [t]#. That is, listeners are basing their judgments on their understanding of the frequency of possible acoustic forms in their language input.

Taken in total, the results reported here suggest that, in fact, speakers do not rely on phonotactics alone when making judgments concerning the goodness of word forms. Certainly, they are aware of the overarching regularities in their language input that make up what we have been calling phonotactic constraints, as seen in their across-the-board preference for canonical forms. But at the same time, they are sensitive to the frequencies of the actual manifestations of a variety of final consonants and final CV syllables. In fact, however, it is not clear how useful the distinction between the implicit understanding of phonotactics and that of possible acoustic forms is. Note that we could as easily say that canonical forms are significantly preferred not because listeners have an implicit understanding of underlying regularities that we call phonotactics, but because these forms occur significantly more frequently in the input than noncanonical forms, just as we did in examining the results of Dupoux and colleagues (Dupoux, Fushimi et al., 1999; Dupoux, Kakehi et al., 1999). Analogously, we might surmise that, as speakers compute frequencies of particular acoustic manifestations of CV sequences in their language input, they form a rule-like understanding of the contexts in which devoicing occurs, which we then call phonotactics. This debate crucially involves the question of the nature and source of representations of native language sound patterns, which is beyond the scope of this paper.

But even laying aside the question of the status of phonotactic representations, we can see that our results clearly indicate that neither source of understanding concerning acceptable word forms is used exclusively. Listeners may base their preference for canonical forms on their implicit understanding of phonotactic constraints, but in the detailed assessment of their perception of word forms made possible by the goodness rating task used here, they show evidence of a fine-grained appreciation of possible word forms as well.

4.3.3
Consecutive devoicing context
Recall that the perception of neeks as a canonical form requires listeners to epenthesize two high vowels in two consecutive contexts of vowel devoicing. Yoshida (to appear) reports devoicing rates of near 0% for the first of two devoiceable vowels, and rates of near 100% for the second vowel; however, both of the vowel contexts in question are word-internal in that study. Maekawa and Kikuchi (to appear) report that in over 25% of consecutive devoicing contexts (both also word-internal), both vowels are devoiced. No data seems to exist for consecutive contexts including a word-final
position. However, it seems clear that forms in which both vowels in these contexts are devoiced would be uncommon.

What are the predictions from phonotactically-oriented and acoustics-based approaches in the case of consecutive devoicing? If phonotactics guide the listeners’ perceptions, results for neekusu should be exactly the same as for keetsu; they are both canonical forms. Likewise, results for neeks should be the same as those for keets; they both depart from canonical forms by the lack of a final vowel. Predictions based on acoustic input are less certain given the lack of production data for consecutive devoicing with a word-final context, but for neekusu and keetsu, it would appear that ratings should be similar since both forms contain only one vowel that can easily be devoiced, and that vowel appears after frication in both forms. In fact, neekusu is rated more highly than keetsu. It is not clear why this should be the case, but then, we know very little about consecutive devoicing in this context; in making our prediction we assumed both that consecutive devoicing is avoided and that, given that avoidance, it is the second vowel that is devoiced. While the first assumption seems valid, the data for the second are less clear.

The case of neeks and keets is clearer. Ratings for neeks should be lower than those for keets, since forms with consecutive vowel devoicing will be less frequent in the input than those with final high vowel devoicing. These are, in fact, the results that we have found, again revealing listeners’ sensitivity to frequencies of acoustic forms in their language input.

5 Summary

How do Japanese listeners perceive devoiced vowels? In voiced contexts in which devoiced vowels never occur in language input, listeners canonicalize the word forms they hear by epenthesizing an appropriate vowel (Dupoux, Fushimi et al., 1999; Dupoux, Kakehi et al., 1999). Word forms that contain contexts in which vowels may be devoiced can be perceived as containing devoiced vowels whether or not the vowels are actually present (Morris, 2004). How do we reconcile these findings?

We tested Japanese listeners’ perceptions of a variety of forms that did and did not contain vowel devoicing contexts, and did and did not contain voiced vowels in those contexts. Listeners’ judgments of the goodness of these forms revealed that they are sensitive to more than a straightforward canonical/noncanonical distinction. Listeners are aware as well of the likelihood of encountering particular types of acoustic forms in their language input, in this case, the likelihood of encountering forms that derive from the legitimate process of vowel devoicing. In words in which vowel devoicing has occurred, listeners prefer those forms that are derived legitimately (keets) to those that are not (keet). Likewise, where voiced vowels are present in the acoustics, listeners prefer voiced vowels that do not occur in a devoicing context (keeto) to those that do (keetsu). Further, listeners have knowledge of the likelihood of encountering forms derived from consecutive devoicing, preferring those that derive from the more common process of single devoicing (keets) to those derived from the less likely process of consecutive devoicing (neeks).
This last point especially argues against a purely phonotactic motivation for these results. It is clear that listeners understand the general constraints that characterize the acoustic forms of their language, constraints that we have called phonotactic, but this understanding could derive from knowledge of the overwhelmingly higher frequency of these types of forms in their language input. That listeners incorporate their understanding of frequencies of acoustic forms into their preference judgments for particular types of forms is clear from the nuances of judgment they exhibit regarding the noncanonical but possible forms in this study.

6 Future directions

The difference between *neeku* and *keetsu* with respect to these results is interesting. If our claim that listeners are sensitive to frequencies of acoustic forms is correct, the answer to why *neeku* patterned as a form with a context that is not conducive to final vowel devoicing may be a straightforward matter of frequency. That is, it may simply be the case that, for whatever reason, [~ku]# forms have a much lower frequency of final vowel devoicing relative to, say, [~tsu]# forms. This is a matter in need of empirical investigation. Indeed, studies specifically examining the nature and frequency of final vowel devoicing in general would shed welcome light on the phenomenon we have investigated here.

In addition, there were large individual differences (SDs) in the ratings that participants gave these forms. We conducted a repeated measures ANOVA on the mean SDs for the eight word types. There was a significant main effect of word type, \( F(7, 273) = 19.813, p < .001 \). In post hoc tests, Fisher’s LSD was used to examine the differences among word types. *Keets, keetsu, neeku,* and *neekusu* did not differ among themselves, but had significantly greater variability than the other items \((p < .05)\). The last three items on this list have voiced vowels in devoiceable contexts (i.e., *keetsu, neeku,* and *neekusu*); this seems to signify that listeners are uncertain about the status of these items. The list includes *keets* as well, which is the form that contains the most likely context for devoiced vowels (as we have seen, *keet* virtually never occurs; *neek* would derive from *neeku,* which is not considered to contain a good context for devoicing by our participants; and *neeks* would have to derive from consecutive devoicing, also avoided). These larger variations are reasonable if, as we have claimed, listeners make their judgments based upon their experience with frequencies of devoicing in their input. The forms for which there are large individual differences in rating, those that have voiced vowels in devoiceable contexts or the one that has no vowel in an eminently devoiceable context, are exactly those for which variable experience in language input has variable impact upon listeners’ judgments. The forms *keet,* *keeto,* and *neeks* are more stable in the sense that they are either very common or very rare in the input ([t]# has very low frequency; [to]# has a high frequency, and [ks]# would have to derive from consecutive devoicing, also quite an uncommon occurrence, leading to a very low frequency).

The apparently problematic case is again the *neek/neeku* case. We have conjectured above that listeners judged *neeku* not to contain a context particularly well-suited to devoicing; this would mean that it should pattern similarly to *keeto.* While its ratings are closer to *keeto* than they are to *keetsu,* *neeku* does not pattern with *keeto*
in terms of how certain listeners are about the status of this form. Listeners are more uncertain about neeku than they are about keeto. This seems reasonable; perhaps the variability in the confidence with which judgments about neeku were made indicates the marginal status of neeku as a form containing a devoicing context. This variability is, again, the hallmark of an acoustic form-based approach to judgments concerning acceptability of word forms.

References


