The goal of the present study is to investigate the degree to which native Japanese speakers’ perception of syllables in spoken English words is affected by various phonetic and phonological factors. Specifically, the present study examined whether Japanese listeners’ perception of English syllables is affected not just by phonological factors such as syllable complexity, but also by acoustic-phonetic factors such as the duration of the spoken English words. In addition, the present study examined the possible influence of another factor, namely, consonant voicing.

Syllables in Japanese are relatively simple, often having a CV (consonant-vowel) or CVC structure, e.g., “ki” (tree) and “hon” (book), but syllables in English tend to have more complex structures such CCCVCC and CVCCC, e.g., “strange” and “text”. Numerous behavioral studies have in fact demonstrated that native speakers of Japanese, when faced with the task of producing or perceiving complex English syllables, often experience difficulties, resulting in mispronunciation or misperception of English syllables (Tarone, 1980; Tajima et al., 2003; Otake et al., 1996; Dupoux et al.,

* This research was funded by a Special Research Grant from Hosei University. This project was also supported in part by a Grant-in-Aid for Scientific Research, No. 16720094, from the Ministry of Education, Culture, Sports, Science and Technology and by the National Institute of Information and Communications Technology. Parts of this research were presented at the 2004 International Congress on Acoustics, April 4-9, 2004, Kyoto, Japan, and at the 2004 fall meeting of the Acoustical Society of Japan, September 28-30, Okinawa, Japan.
In particular, a series of studies using a syllable-counting task have demonstrated the following results: (1) Native Japanese listeners’ ability to count syllables in spoken English words is strongly affected by syllable complexity of the stimuli, as measured by the number of consonants in the item (Tajima and Erickson, 2001). (2) Listeners’ performance is also affected by whether certain consonants were voiced or voiceless, e.g., /peps/ vs. /pebz/ (Tajima and Erickson, 2001). (3) Perceptual identification training using feedback significantly improves native Japanese listeners’ performance, even after just one or two hours of training (Tajima and Erickson, 2001). (4) The effect of training generalizes to untrained words, and is retained several months after the completion of training (Tajima and Akahane-Yamada, 2003). (5) Training is effective not just for college-aged Japanese listeners, but also for elderly Japanese listeners in their 60’s (Tajima et al., 2002).

Previous studies have repeatedly indicated a strong effect of syllable complexity; that is, as syllable complexity of the stimulus increases, the ability to identify the number of syllables decreases (Tajima and Erickson, 2001; Tajima et al., 2002). For example, identification accuracy was close to 100% for nonword stimuli that started with no initial consonants, e.g., /ep/, but was approximately 40% for nonwords that started with a three-consonant cluster, e.g., /splep/. This result has been interpreted to mean that the relatively simple syllable structure of Japanese, along with the role of moras in Japanese rhythm, lead native Japanese listeners to perceive multiple prosodic units in complex English syllables.

However, English words with complex syllable structure tend to have longer word duration at the same time. Therefore, an alternative interpretation of the result is possible, namely, that Japanese listeners based their judgment to some extent on “how long” a stimulus was (a phonetic factor) rather than on “how many units” it contained (a phonological factor). Because word duration covaried with syllable complexity in previous studies, the possibility that listeners relied on absolute duration when counting syllables cannot be ruled out. Given that duration serves as an important cue for the perception of Japanese moras (e.g., Kato et al., 2004), it is possible that Japanese listeners might show sensitivity to durational cues when perceiving prosodic units in
English as well. To tease apart the effect of syllable complexity and acoustic duration, the present study used English words produced at multiple speaking rates as stimuli in a syllable counting task.

A close investigation of the effect of speaking rate is warranted for several reasons. First, speech perception in general is known to be highly sensitive to contextual information such as differences in speaking rate (e.g., Summerfield, 1981). For example, the phoneme boundary along a /ba/-/pa/ VOT continuum could be shifted by changing the speaking rate of the context in which the continuum occurs. Second, it has been claimed in the second-language (L2) speech learning literature that L2 learners have difficulty exploiting contextual information when perceiving L2 utterances, and that L2 learners would have greater difficulty than native listeners in adapting to changes in speaking rate (Toda, 2003). For example, native English listeners’ identification of Japanese words minimally contrasting in phonemic length, e.g., /kado/ (corner) vs. /ka:do/ (card), was relative good for words produced at a normal rate, but poorer for words produced at slow or fast rates (Tajima et al., 2002) . Furthermore, in the same study, when compared to a fixed-rate condition in which the speaking rate of the stimuli was fixed, native English listeners’ identification accuracy was lower in a mixed-rate condition in which the speaking rate varied from trial to trial. Given these results, it is reasonable to investigate whether effects of speaking rate are observed for native Japanese listeners’ perception of English syllables.

In addition, a previous study has also found a modest effect of consonant voicing (Tajima and Erickson, 2001). For example, mean accuracy for words that end with a voiceless consonant cluster, e.g., /peps/, was 45%, while accuracy for words that end with a voiced consonant cluster, e.g., /pebz/, was 38%, which was significantly lower. Several interpretations of this result are possible: (1) Given that vowel duration generally tends to be longer before voiced consonants than before voiceless consonants, Japanese listeners may have shown a tendency to count more units in items ending in voiced consonants (which are preceded by longer vowels on average) than voiceless consonants (which are preceded by shorter vowels). (2) When Japanese speakers produce English words, they tend to insert epenthetic vowels, i.e., extra vowels that are
not part of the canonical pronunciation of the word, e.g., “sneezed” pronounced as /sunizudo/; furthermore, they show a greater tendency to produce epenthetic vowels between or following voiced consonants than voiceless consonants (Tajima et al., 2003). If it is assumed that listeners count syllables based on their own productions of the words, then they are more likely to overestimate the number of syllables in words containing voiced consonants (which tend to contain more epenthetic vowels and hence more syllables) than in words containing voiceless consonants. In other words, given the asymmetry between voiced and voiceless environments in syllable production, a similar effect might be observed in syllable perception as well, assuming that listeners make reference to their own production tendencies while they perform perception tasks. To examine which of the above interpretations might be more feasible, a syllable counting experiment was conducted with nonword stimuli contrasting in voicing not just in post-vocalic position but also in pre-vocalic position. In addition, the voicing contrast occurred either in a single consonant or in a consonant cluster.

2. Method

2.1. Stimulus materials

Stimulus materials consisted of 40 English words of 1-6 syllables in length and 44 nonwords made up of 22 monosyllabic and 22 disyllabic items. The 40 English words were a subset of the words used in previous tests (see Tajima, 2003, Appendix, for a list of the test words). Of the 44 nonwords, 32 were the same as those used in previous studies to test the effect of syllable complexity. In these nonwords, the number of initial and final consonants each varied from zero to three in a factorial manner (initial consonants = {none, /p-/, /sp-/, /spl-}, final consonants = {none, /-p/, /-ps/, /-mps/}) (see Tajima, 2003 for further details). To these nonwords were added six monosyllabic and six disyllabic nonwords that contained voiced rather than voiceless consonants in either initial or final position or both: monosyllabic nonwords = /pebz/, /beps/, /bebz/, /speb/, /blep/, /bleb/, disyllabic nonwords = /pedezb/, /bedeps/, /bedebz/, /spedeb/, /bledep/, /bledeeb/.
Each word and nonword was produced at three self-selected speaking rates (fast, normal, slow) by a male native American English talker (aged 33). Recordings were made in an anechoic chamber at ATR Laboratories, Kyoto, Japan, and were digitized at 22.05-Hz sampling frequency and 16-bit resolution. Individual words were amplitude-normalized so that the peak amplitude was constant across all items.

A group of 21 native Japanese college students at Hosei University (9 females, 12 males, aged 18-24, mean age = 20.2 with no prior experience living in an English-speaking community for over three months participated in the study.

Prior to the experiment, participants read a brief textual description of English syllables, and performed a block of 10 practice trials using a different set of words and a different talker from the test. The practice trials were identical to the test trials except that participants received immediate feedback about their response (correct or incorrect), and repeated a given trial until the correct response was entered.

On each test trial, participants heard a stimulus through headphones at a comfortable listening level and counted the number of syllables in the stimulus by clicking one of ten buttons numbered “1”–“10” in the program window. Items produced at the three rates were all presented within a single block, so that speaking rate varied from trial to trial. Each item was presented once, for a total of 252 trials (84 items X 3 rates).

Acoustic duration of each nonword stimulus (44 stimuli X 3 rates = 132 tokens) was measured based on waveform and spectrogram displays of each token. These measurements were pooled and averaged in order to examine how word duration varied as a function of speaking rate and syllable complexity.

The 21 listeners’ responses to the 252 tokens (yielding 5292 observations) were used to compute mean percent-correct identification accuracies as a function of speaking rate, syllable complexity, and consonant voicing. To test for statistical significance, arcsine-transformed values of the percent-correct scores were submitted to
analyses of variance (ANOVA) with listener as a random factor and various subsets of the three factors as within-subjects variables.

3. Method

3.1. Speech Durations

Figure 1 shows the mean duration of the 40 English words and 44 nonwords as a function of speaking rate. The figure clearly shows that mean word duration increased as speaking rate decreased from fast to normal to slow. For both real words and nonwords, the mean duration ratio of items produced at the fast and slow rates was 1:1.65. The longer mean duration of the real words compared to the nonwords stems from the fact that the real words were of 1-6 syllables in length, while the nonwords were of 1-2 syllables in length.

Turning to the effect of syllable complexity on word duration, there was a highly consistent effect of syllable complexity on word duration. As syllable complexity of the nonwords increased, as measured by the total number of word-initial and word-final
Phonetic and phonological factors in the perception of syllables in second-language speech

Consonants, word duration increased monotonically, starting from a mean duration of 439 ms for nonwords containing no initial or final consonants, to a mean duration of 890 ms for nonwords containing a total of six initial and final consonants.

Figure 2 plots mean word duration as a function of the number of word-initial consonants. Separate means are shown for words produced at the three speaking rates. The figure shows an overall trend for word duration to increase as the number of initial consonants increases, but the greatest increase is observed between words starting with /p-/ and words starting with /sp-/; presumably stemming from the long inherent duration of /s/.

Figure 3 plots mean word duration as a function of the number of word-final consonants as well as speaking rate. Here again, there appears to be an overall trend for word duration to increase as the number of final consonants increases, but the greatest increase is observed between words ending in /-p/ and words ending in /-ps/, again possibly stemming from the long inherent duration of /s/.
Figure 4 plots the Japanese listeners’ mean identification accuracy for the words and nonwords as a function of speaking rate. The figure indicates that accuracy is more or less constant across the three speaking rates, in spite of the fact that word duration varied by a ratio of 1:1.65 between the fastest and slowest rates (compare Figure 4 with Figure 1). For real words, mean accuracy was between 49% and 53%, and for nonwords, mean accuracy was between 61% and 63%. In short, it appears that speaking rate had negligible effects on the native Japanese listeners’ identification performance.

A comparison was further made between results from the present experiment which employed a mixed-rate design, and a previous study which used a fixed-rate design. The mean accuracies for words and nonwords in the present study were 51% and 62%, respectively, while the mean accuracies for words and nonwords in Tajima et al.,’s (2002) study were 51% and 60%, respectively. Judging from the similarity of the accuracies between the two studies, it appears that trial-by-trial variation in speaking

(1) Real words showed lower accuracies than did nonwords presumably because the real words showed greater variability in segmental content and in length (1~6 syllables) than the nonwords (1~2 syllables).
rate does not increase the difficulty of the syllable counting task for Japanese listeners.

Turning to the effect of syllable complexity, there was a highly consistent effect of syllable complexity on identification accuracy, as has been reported previously. As syllable complexity of the nonwords increased, accuracy decreased monotonically, starting from a mean accuracy of 88% for nonwords with no initial or final consonants, to a mean accuracy of 40% for nonwords with six initial and final consonants.

Figure 5 plots mean identification accuracy as a function of the number of initial consonants and speaking rate. An ANOVA with rate (fast, normal, slow) and initial consonant (0~3) as within-subjects factors revealed a significant main effect of initial consonant \( [F(3,60) = 57.96; \ p < .01] \). No other main effects or interactions were significant. When the main effect of initial consonant was further analyzed by means of post hoc multiple comparisons by LSD, results revealed the following pattern of significant differences at the \( p < .05 \) level among the four levels of the factor: 0 = 1 > 2 > 3. That is, there were no significant differences in accuracy between nonwords that begin with no initial consonant and nonwords that begin with /p-/ but accuracy significantly dropped as the initial consonants increased in complexity from /p-/ to /sp-/ to /spl-/.

The largest drop in accuracy is observed between nonwords starting with /p-/ and those
starting with /sp-/ suggests that English word-initial clusters containing /s/ pose particular difficulties for native Japanese listeners.

Speaking rate does not seem to have a consistent effect on accuracy in Figure 5. This is so in spite of the variation in word duration across speaking rates, as depicted in Figure 2. For words that begin with /sp-/ and /spl-/ there appears to be a slight tendency for accuracy to decrease as speaking rate increases.

Figure 6 plots mean accuracy as a function of the number of final consonants and speaking rate. An ANOVA with rate and final consonant as within-subjects factors revealed a significant main effect of final consonant \([F(3,60) = 14.17; p < .01]\). No other main effects or interactions were significant. Post hoc comparisons for the main effect of final consonant revealed significant differences among all possible pairs of accuracies, except for the comparison between /-p/ and /-ps/ and between /-ps/ and /-mps/. In other words, accuracy significantly declined as the final consonants increased in complexity, but the decline was somewhat smaller for stimuli with a greater number of final consonants.

Speaking rate again does not seem to exert an influence on accuracy, despite the
Phonetic and phonological factors in the perception of syllables in second-language speech

Finally, turning to the effect of consonant voicing, the left half of Figure 7 plots mean accuracy as a function of voicing for nonwords that begin with singleton conso-

variation in word duration observed in Figure 3.
nants (/p-/ vs. /b-/) and for nonwords that begin with a 2-consonant cluster (/sp-/ vs. /bl-/). One-way ANOVAs carried out separately on the singleton and cluster data with voicing (voiceless, voiced) as a within-subjects factor revealed no significant effects. In other words, Japanese listeners’ accuracy did not differ systematically depending on whether the nonwords started with /p-/ or /b-/ or whether they started with /sp-/ or /bl-/. 

In the right half of Figure 7, mean accuracy is plotted as a function of voicing for nonwords that end with singleton consonants (/ -p/ vs. /-b/) and for nonwords that end with a 2-consonant cluster (/ -ps/ vs. /-bz/). A one-way ANOVA carried out on the singleton data indicated no significant effect of voicing. However, a one-way ANOVA carried out on the cluster data revealed a marginally significant effect of voicing [F(1,20) = 3.74; p = .067]. That is, while accuracy did not vary between nonwords ending in /-p/ and nonwords ending in /-b/, there was a small trend for accuracy to be lower for nonwords ending in /-bz/ than for nonwords ending in /-ps/.

Together, these results suggest that the effect of voicing is restricted to word-final consonant clusters, and does not occur with word-final singleton consonants nor with word-initial consonants. The fact that a voicing difference in word-final singleton consonants did not affect accuracy suggests that Japanese listeners were not paying attention to voicing-induced vowel duration differences while performing the task. If they had been paying attention to vowel duration, then a voicing effect should have been obtained for both word-final singletons and clusters. A better candidate to explain the voicing effect might be that listeners are, to a small extent, sensitive to the likelihood of vowel epenthesis in their own productions of English words.

4. Conclusion and Discussion

To conclude, this study had two main purposes: (1) to test whether Japanese listeners’ perception of English syllables is affected not just by syllable complexity (a phonological factor), but also by word duration (a phonetic factor), and (2) to examine the extent to which consonant voicing affects identification performance. Results from a syllable counting experiment with English word and nonword stimuli spoken at three
speaking rates indicated the following: (1) Syllable complexity has a substantial effect on performance, as has been demonstrated in previous studies. (2) In contrast, word duration has virtually no effect on identification accuracy. (3) Similarly, consonant voicing has only limited effects on accuracy.

As Figures 2 and 3 show, word duration covaried substantially with syllable complexity. Because of this, it had been difficult in previous studies to determine whether the apparent relationship observed between syllable complexity and accuracy (compare Figures 2 and 5 and Figures 3 and 6) was to be attributed to the number of consonants or to word duration. However, when variation in syllable complexity was canceled out and only variation in word duration remained, it was found that accuracy was virtually constant (compare Figures 1 and 4), strongly suggesting that word duration by itself does not affect accuracy.

In addition to the absence of an effect of word duration, the trial-by-trial variation in speaking rate in the current study did not lead to a decline in performance compared to conditions in which the rate was fixed. These findings suggest that Japanese listeners are more or less able to ignore variation in speaking rate when judging the number of syllables in spoken English words. Even though native Japanese listeners exploit durational cues in the perception of moras in Japanese (Kato et al., 2004), they do not rely on duration in the perception of English syllables. From a practical standpoint, then, it appears that variation in speaking rate is not a crucial factor to be considered when designing effective training programs for perception of English rhythm.

This situation appears to be quite different from that observed in the perception of phonemic length contrasts in Japanese by native English listeners. That is, native English listeners’ ability to identify words among minimal pairs such as “kado” vs. “ka:do” was better for stimuli produced at certain rates and worse for others, and furthermore, it was worse in a mixed-rate condition than in a fixed-rate condition (Tajima et al., 2003). In addition, it has been suggested that perceptual adjustment to variation in speaking rate or temporal context is difficult for L2 learners, and should therefore be considered when designing training programs (Toda, 2003). However, results from the present study suggest that while there are cases where adaptation to
speaking rate variation is problematic for L2 learners, it is not always the case. Whether adaptation is difficult or not appears to depend on the task for the listeners. In fact, contrary to Toda’s (2003) findings, it has been shown that with extensive perceptual training, native English listeners can learn to shift their perceptual category boundaries for Japanese phonemic length contrasts in accordance with changes in temporal context (Kato et al., 2004). Further work is necessary to determine the nature of L2 learners’ adaptation abilities and the extent to which they approach those of native listeners.

As for the effect of consonant voicing, it appears that the effect of voicing that was found in a previous study (Tajima and Erickson, 2001) is only limited to word-final consonant clusters, and does not generalize to word-final singleton consonants, initial singleton consonants, or initial consonant clusters. Even for word-final consonant clusters, the size of the effect is quite small, especially when compared to the effect of syllable complexity.

Such a limited effect of voicing in syllable perception is in contrast to the effect that consonant voicing has on English syllable production. Studies on vowel epenthesis have shown that Japanese speakers are more likely to produce epenthetic vowels in voiced than voiceless consonant environments. If vowel epenthesis is taken as a diagnostic of difficulties in English syllable production, then it appears that production and perception difficulties do not match straightforwardly. That is, what may be difficult in production may not necessarily correspond to what may be difficult in perception, suggesting only a weak link at best between speech production and speech perception.

Taken together, these results suggest that, among many factors examined thus far, syllable complexity is a principal determinant of the degree to which Japanese listeners can accurately count syllables in spoken English words. The precise mechanisms that underlie the Japanese listeners’ performance are not yet clear, and require further investigation.
Phonetic and phonological factors in the perception of syllables in second-language speech


