Epenthetic Vowels in Japanese: A Perceptual Illusion?

Emmanuel Dupoux
Ecole des Hautes Etudes en Sciences Sociales–Centre National Pour la Recherche Scientifique

Yuki Hirose
City University of New York

Kazuhiko Kakehi
Nagoya University

Christophe Pallier and Jacques Mehler
Ecole des Hautes Etudes en Sciences Sociales–Centre National Pour la Recherche Scientifique

In 4 cross-linguistic experiments comparing French and Japanese listeners, we found that the phonotactic properties of Japanese (a reduced set of syllable types) induce Japanese listeners to perceive “illusory” vowels inside consonant clusters in vowel-consonant-consonant-vowel (VCCV) stimuli. In Experiments 1 and 2, a continuum of stimuli ranging from no vowel (e.g., /ebu/ko/) to a full vowel between the consonants (e.g., /ebu/ko/) was used. Japanese, but not French participants, reported the presence of a vowel [u] between consonants, even in stimuli with no vowel. A speeded ABX discrimination paradigm was used in Experiments 3 and 4 and revealed that Japanese participants had trouble discriminating between VCCV and VCVCV stimuli. French participants, in contrast, had problems discriminating items that differed in vowel length (/ebu/ko vs. /ebu/ko/), a distinctive contrast in Japanese but not in French. It is concluded that models of speech perception have to be revised to account for phonotactically based assimilations.

Human languages differ in the sound contrasts used to distinguish words. A contrast between two phones (e.g., a bilabial voiced stop and a bilabial unvoiced stop) may signal a difference in meaning in one language but not in another. It has been known for a long time that this has an influence on the perception of speech sounds: Speakers of a given language often have trouble distinguishing nondistinctive phones (e.g., Sapir, 1921). For example, Japanese listeners map the American [l] and [r] phones onto their own single [R] category and therefore have trouble discriminating them. However, not all foreign contrasts are difficult: In fact, they vary in the degree of perceptual difficulty (Best, McRoberts, & Sithole, 1988; Polka, 1991). It has been only recently that the perception of nonnative speech sounds has been systematically explored and that theories to account for it have been advanced. For instance, in Best’s (1994) perceptual assimilation model, a foreign sound can be processed in one of two ways. If the phonetic characteristics of that sound are close to those of an existing phoneme category in the maternal language, the sound will be assimilated into that category. In this case, listeners will be able to judge only whether it is a good or a bad exemplar of that category but will not have access to its detailed phonetic characteristics. (In particular, two equally bad examples but phonetically distinct of a category will be difficult to discriminate.) In contrast, if the foreign sound is too distant from any of the available categories, it will not be assimilated at all and listeners will have conscious access to its fine phonetic characteristics.

The perceptual assimilation model is meant to account only for the effects of the phonemic repertoire. However, human languages also differ in the rules that govern what sequences of phonemes are allowed in an utterance. For instance, some languages (e.g., French or English) allow complex clusters of consonants, whereas others (e.g., Japanese) disallow them. One may expect that language-specific constraints play a role in speech perception and that language-specific influences may be demonstrated that go beyond phonemic categorization. For instance, in Spanish, /s/-/r/-consonant clusters are always preceded by a vowel and we have informally heard reports by Spanish speakers who maintain that they hear the vowel [e] preceding English words that begin with an /s/ cluster. Accordingly, many Spanish speakers of English sometimes produce especial instead of special, estimulus instead of stimulus, esport instead of sport, and so on. This has nothing to do with the phonemic categories of [s] and [e] in Spanish versus English; instead, it seems to depend on a Spanish-specific phonotactic property.

Emmanuel Dupoux, Christophe Pallier, and Jacques Mehler, Ecole des Hautes Etudes en Sciences Sociales–Centre National Pour la Recherche Scientifique, Paris, France; Kazuhiko Kakehi, Cognitive Informatics Unit, Nagoya University, Nagoya, Japan; Yuki Hirose, Department of Linguistics, City University of New York.

This work was supported by grants from the Fyssen Foundation, the Human Frontiers Science Program, the Human Capital and Mobility Program, and the Direction de la Recherche et des Techniques.

We thank Dianne Bradley, Susana Franck, Takao Fushimi, Peter Golato, Takashi Otake, and Sharon Peperkamp for useful comments on this article and discussion. We thank Stanka Fitneva, Olivier Crouzet, and Laurent Somek for preparation and running of the experiments. We are especially grateful to Hideo Yamashiki for invaluable help in recruiting Japanese participants and Alain Grumbach for providing access to French participants. We also thank Franck Ramus and Evelyn Hausklein for additional help in recruiting participants.

Correspondence concerning this article should be addressed to Emmanuel Dupoux, 54 Bd Raspail, 75006 Paris, France. Electronic mail may be sent to dupoux@lscp.ens.fr.
In this article, we focus on a similar phenomenon in Japanese. As we indicated earlier, the Japanese language disallows complex consonant clusters. This is a consequence of the language's simple syllable structure: Indeed, the complete syllable inventory of Japanese consists of vowel (V), CV, consonant-vowel (CV), consonant-vowel-nasal (CVN), and CVQ (where Q is the first half of a germinate consonant). This can be illustrated by loan words, that is, words of foreign origin that were changed to conform to the Japanese pattern (see below; Ito & Mester, 1995):

- **fight** → faiito
- **festival** → fesutibaru
- **sphinx** → sufiNkusu
- **Zeitgeist** → tsaitogaisuto

As we can see, [u] or [o] is inserted after every syllable-final consonant (except for nasal consonants). Why do the Japanese insert vowels in loan words? A first possibility is that this phenomenon (called vowel epenthesis) arises in speech production. Perhaps Japanese speakers have, to some extent, lost or fail to develop the ability to articulate consonant clusters and therefore tend to insert vowels to trigger the more practiced CV motor programs. A second possibility is orthography: Kanji orthographic characters, by and large, are pronounced as either [n], V, or CV. Hence, there is no Japanese character or combination of characters (in the kanji system) that can spell an item like /sinks/ or any other item with a consonant cluster that does not include nasals. In contrast, /sufinkusu/ can easily be spelled in Japanese. Could it be that Japanese speakers modify foreign words by inserting epenthetic vowels so that they can be spelled in their language? In this article, we explore a third possibility: according to which vowel epenthesis will occur at the perceptual level. Assessing the perceptual reality of epenthesis is important because it bears on the role of phonotactics in speech perception.

What evidence exists that phonotactic constraints play a role in perception? Adults have clearer intuitions about permissible sequences. For example, English speakers know that “mba” is not a possible English word. Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk (1993) and Jusczyk, Luce, and Charles-Luce (1994) have shown that 9-month-old infants are sensitive to the phonotactic patterns of the words in their language, and some researchers have argued that such regularities could be useful in helping the child to discover words (Brent & Cartwright, 1996; Hayes & Clark, 1970). Massaro and Cohen (1983) investigated the influence of phonotactic constraints on phoneme perception. They used the fact that /sri/ and /fi/ are not allowed in English but that /sl/ and /fr/ are allowed. They synthesized a series of stimuli ranging from [s] to [f] and presented them to participants in the /sl/ and /fr/ context. There was a significant shift in the identification functions between the two contexts, demonstrating that participants tend to hear segments that respect the phonotactics of their language. Notice, however, that Massaro and Cohen's study (1983) demonstrates only an effect on ambiguous stimuli. It would be desirable to demonstrate the influence of phonotactics on end point (unambiguous) stimuli. Second, their study was conducted with a single language, leaving open the possibility that some of the effects might be found in all speakers regardless of their native language. Hallé, Segui, Frauenfelder, and Meunier (1998), using natural stimuli and various tasks, showed that illegal French syllables such as /diла/ tend to be assimilated to legal ones, such as /gli/ Again, that study was conducted within a single language, leaving open the possibility that part of the observed effect might have been due to universal properties of phonetic perception.

Here, we further explored the role of phonotactics on perception using a methodology that involved nondegraded speech stimuli and a cross-linguistic design. We investigated the perceptual reality of epenthesis using an off-line phoneme detection task (Experiments 1 and 2) and two speeded ABX tasks (Experiments 3 and 4). We tested the same stimuli on two populations: native Japanese speakers and native French speakers. French has complex syllabic structures and hence should not trigger epenthetic effects. Comparing the performances of French and Japanese participants on exactly the same materials allowed us to assess how language experience would influence the perception of these stimuli.

**Experiment 1**

The aim of this experiment was to assess the extent of the epenthesis effect. We created nonword stimuli that formed a continuum ranging from trisyllabic tokens such as ebuço to disyllabic tokens such as ebo to progressively removing acoustic correlates of the vowel from the original stimuli. We selected our materials so that the word internal consonant clusters would always yield an epenthetic [u] in Japanese (e.g., the first consonant of the cluster was not a nasal or a dental stop). French and Japanese participants were then asked to decide whether the vowel [u] was present in the stimuli. No overt production of the stimuli was needed. If the epenthesis effect has a perceptual basis, Japanese participants should report the presence of [u] more often than French listeners.

**Method**

**Participants.** Ten Japanese and 10 French native speakers volunteered to participate in Experiment 1. All the participants were college students. The French participants were recruited in Paris and the Japanese at Nagoya University. None of the Japanese

---

1. The inserted vowel is most often [u], except after a dental stop, in which case it is an [o] (see Shinohara, 1997, for a more complete discussion).
2. McClelland and Elman (1986) claimed that such effects are not due to phonotactics per se but to top-down word to phoneme activation during perception (see Massaro & Cohen, 1991, for a reply).
participants had studied French, and none of the French participants had studied Japanese.

Materials. Ten sequences of VC1uCV (V = four Japanese vowels excluding [u], C1 = voiced and voiceless stops, and C2 = nasals and voiceless obstruents) uttered by a male Japanese speaker were used as the stimulus items (see the Appendix). None of the stimulus items constituted a meaningful word in French or in Japanese.

The stimuli were digitized at 16 kHz/16 bits on a PC-compatible computer using an OROS AU22 A/D board. Five different files were then created from each original item by splicing out pitch periods of the medial vowel [u] at zero crossings. Stimulus 1 contained little or no vowel [u] (most of the transitions in and out of the vowel were also removed). Stimulus 2 contained the two most extreme pitch periods of the vowel [u], one from the transition of the first consonant to the vowel [u] and another from the end of the vowel [u] into the following consonant. Stimulus 3 contained the four most extreme pitch periods (two on each side); similarly, Stimulus 4 contained six pitch periods and Stimulus 5 contained eight pitch periods. Stimulus 6 was the original stimulus in which the number of pitch periods varied from 10 to 13 across items (10.7 periods on average). The average overall duration of one pitch period in the [u] vowels in each item was 9.06 ms. There were a total of 60 stimuli in one session.

Procedure. Participants were instructed to listen to the stimuli through headphones and to judge whether there was a [u] vowel in

| Figure 1. Percentage (y axis) of [u] vowel judgments in stimuli such as *ebu*o in French and Japanese participants as a function of vowel duration (x axis) in Experiment 1. |

The two languages have significantly different vowel lengths (0, 8, 16, and 24 ms). The vowel length in Japanese is 16 ms, whereas the vowel length in French is 24 ms. The interaction of vowel length and pitch period was significant, F(7, 98) = 2.36, p < .05.

The two languages have significantly different vowel lengths (0, 8, 16, and 24 ms). The vowel length in Japanese is 16 ms, whereas the vowel length in French is 24 ms. The interaction of vowel length and pitch period was significant, F(7, 98) = 2.36, p < .05.

In the French participants, the percentage of correct responses was significantly lower than in the Japanese participants. The vowel length and the pitch period were significant factors, F(7, 98) = 3.24, p < .01.

In the Japanese participants, the percentage of correct responses was significantly lower than in the French participants. The vowel length and the pitch period were significant factors, F(7, 98) = 3.24, p < .01.

Discussion

In this experiment, Japanese and French participants judged the presence or absence of the vowel, containing varying extents of the acoustic continuum of the vowel. French participants were able to judge the presence of the vowel, as shown by the significant main effect of vowel length. However, the vowel length effect was not as strong as the vowel duration effect, F(1, 26) = 5.25, p = .03. In the Japanese participants, the vowel duration effect was significant, F(1, 26) = 5.25, p = .03. The vowel duration effect was not as strong as the vowel length effect, F(1, 26) = 5.25, p = .03.

The mean percentages of vowel responses as a function of language and vowel length are shown in Figure 1. We performed two analyses of variance (ANOVAs) on percentages of vowel responses, one with participants and one with items as random variables. Language (Japanese or French) was a between-subjects factor and vowel length, a withinsubjects factor (with six levels). In the first analysis, we report the Min F statistics when they were significant (p < .05) and the F1 and F2 statistics otherwise.

Overall, there was a significant language effect, Min F(1, 25) = 25.10, p < .001, with the Japanese participants providing more vowel responses than the French participants. There was also a significant vowel length effect, Min F(5, 100) = 5.18, p < .001, which had a significant linear component, Min F(1, 20) = 152.32, p < .001. In that longer vowels yielded more vowel responses than shorter vowels. There was an interaction between language and the linear component of vowel length, Min F(1, 26) = 128.62, p < .001, corresponding to the fact that the French participants were much more influenced by vowel length

3 The linear component is defined as a zero-order coefficient applied across the six levels of vowel length, Winer, Brown, & Michels, 1991, pp. 148, 198–240.
been removed, they still reported that the vowel was present in more than 70% of the cases.

At this point, we raise the following caveat. Even though we digitally removed the vocalic [u] portions of VCuCV stimuli, it is unlikely that it was completely deleted. Indeed, coarticulatory information for the rounded vowel was likely to be present throughout the portion of acoustic signal for the surrounding consonants. It may be that unlike French, Japanese listeners have developed a fine capacity for perceiving short vowels. One reason might be that in Japanese, the

\[\text{[u]} \Rightarrow \text{VCuCV occasionally gets devoiced (Beckman, 1952, Keating & Hoffman, 1984).} \]

Another reason is that the speaker that we used in Experiment 1 was Japanese, and he therefore might have introduced coarticulation cues in the adjacent consonants that were especially salient to Japanese hearers.

Therefore, even though we identified a perceptual difference between members of two language communities, the difference might not be due to phonotactics. The next experiment was designed to address this issue.

**Experiment 2**

In Experiment 2, we used a paradigm similar to the one used in Experiment 1 with the following modifications: We recorded a French speaker and digitally generated similar continua to those used in the previous experiment (ebu-zo-ebzo). In addition, we recorded two extra conditions: one with no vowel, that is, a naturally produced consonant cluster (ebzo), and one condition with a vowel different from [u] (ebzo). This last condition was introduced to measure baseline performance.

If the results of Experiment 1 were due to coarticulation information about the vowel on the adjacent consonants, then we should expect that Japanese [u] responses on the naturally produced ebzo should drop to the baseline level. If phonotactics, then naturally-produced clusters (ebzo) should produce at least as many [u] responses as the artificially produced clusters.

**Method**

**Participants.** Ten Japanese and 10 French native speakers volunteered to participate in Experiment 2. All participants were college students. The French participants were recruited in Paris and the Japanese at Nagoya University. None of the Japanese participants had studied French, and none of the French participants had studied Japanese.

**Materials.** We used the same 16 sequences of VCuCV stimuli as in Experiment 1. To these stimuli, we added 10 corresponding VCuCV stimuli (with VCuCV stimuli with VCuCV stimuli different from [u] and from VCuCV). The stimuli were uttered by a male French speaker. None of the stimulus items constituted a meaningful word in French or in Japanese.

The stimuli were digitized at 16kHz/16 bits on a PC-compatible computer using an OROS AT/22 A/D board. As in Experiment 1, we digitally generated five extra stimuli from the VCuCV stimuli by splicing out pitch periods of the medial vowel [u] at zero crossings. Stimulus 1 contained little or no vowel [u] (most of the transitions in and out of the vowel were also removed). Stimulus 2 contained the two most extreme pitch periods of the vowel (i.e., one from the transition of the first consonant to the vowel [u] and another from the end part of [u] into the following consonant). Stimulus 3 contained the four most extreme pitch periods (two on each side); similarly, Stimulus 4 contained six pitch periods and Stimulus 5 eight pitch periods. Stimulus 6 was the original production.

**Procedure.** Participants were instructed to listen to the stimuli through headphones and to judge whether a [u] vowel was present in the middle of each stimulus word. The whole stimulus set was played three times, each time in a different pseudorandom order from a PC-compatible computer with a ProAudio spectrophonic digital-to-analog board using the EXPE program (Pelletier, 1992; Jeannin, 1997). The participants had to press one key if the vowel was present and another if it was absent. Otherwise, the same procedure as in Experiment 1 was used.

**Results**

The mean percentage of vowel judgments of vowel length are shown in Figure 2. We performed two sets of analyses. The first was identical to that used in Experiment 1 and analyzes the effect of the six VCuCV stimuli in the second set of analyses of coarticulation more directly and VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli. In the second set of analyses of coarticulation more directly and VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli of VCuCV stimuli.

On consonant length, MinF '(1, 126) = 4.97, p < .0001, which yielded a significant difference between the two languages, with Japanese participants responding to the longer VCuCV stimuli. There was an interaction between language and the component of vowel length, MinF '(1, 25) = 15.48,
p < .001, corresponding to the fact that the French participants were much more influenced by vowel length than the Japanese. However, even in the Japanese participants, the linear component of vowel length was significant, MinF'(1, 12) = 10.19, p < .008.

We ran pairwise comparisons between the two languages for each vowel length. For the first three vowel lengths (0, 15, and 29 ms), Japanese participants gave significantly more vowel responses than French participants (all MinF' Bonferroni corrected p < .02). For the fourth vowel length (44 ms), there was only a trend in the same direction (Bonferroni-corrected p = .10). A significant difference between the two populations did not appear for the last two vowel lengths (58 ms and full vowel).

Effect of coarticulation. We performed two ANOVAs on the percentages of vowel responses, one with participants and one with items as random variables. Language (Japanese or French) was a between-subjects factor and stimulus type, a within-subjects factor (with three levels: natural cluster, digital cluster, and different vowel).

We found an overall effect of language, MinF'(1, 25) = 19.77, p < .001, stimulus type (MinF'(2, 53) = 17.99, p < .001, and an interaction between them, MinF'(2, 54) = 13.34, p < .001. Individual post hoc contrasts revealed that this interaction was due to the fact that for the French participants, the natural cluster condition was not different from the baseline condition (Fs < 1), whereas the digital cluster condition elicited slightly but significantly more [u] responses than either baseline or natural clusters (ps < .03). In contrast, for the Japanese listeners, stimuli in both the natural and digital clusters condition elicited considerably and significantly more [u] responses than did baseline stimuli (ps < .0001), and the two kinds of clusters did not differ from each other (Fs < 1).

Discussion

In this experiment, we replicated the pattern found in Experiment 1. Moreover, we found that this pattern of results could not be attributed to coarticulatory cues left in the original Japanese tokens. We used tokens produced by a French speaker and compared digitally produced clusters (which might have residual coarticulation information) and naturally produced clusters (which have no coarticulatory information for a vowel). We found that Japanese participants did not perceive more [u] vowels in digital than in natural clusters (in fact, there was a nonsignificant trend in the other direction). Hence, we observed that a majority of [u] responses did arise in Japanese participants even in the total absence of [u] information in the signal.

The results of Experiments 1 and 2 establish that, in a task involving no overt speech production, Japanese participants consistently report a vowel between two consonants in CC clusters. These experiments alone, however, cannot firmly establish the perceptual locus of the effect for two reasons. First, the task required participants to make an explicit metalinguistic judgment: Participants had to know what a vowel is to do the task. It is known that learning to read influences the way in which individual phonemic segments can be manipulated in a metalinguistic task (see Bertelson, 1986). Given that the writing systems of Japanese and French differ, it is possible that they differentially affect vowel judgments in Japanese and French participants. Second, the task did not use a speeded or on-line judgment. Therefore, it could not identify which of the different sources of information (e.g., the orthographic code, covert production, explicit strategies) influenced the participants' responses. For instance, it is possible that Japanese participants were reluctant to give a vowel-absent response simply because they knew that such stimuli do not occur in Japanese.

In the next two experiments, we used an ABX paradigm that required only identity judgments, thus involving no explicit or implicit mention of vowels. We also had participants perform a speeded response, thereby reducing the likelihood of them using complicated response strategies.

Experiment 3

In this experiment we used a speeded ABX paradigm in which participants heard three stimuli in a row and had to decide whether the third stimulus was the same as the first or the second. If the findings of Experiments 1 and 2 have no perceptual basis but are instead a by-product of metalinguistic limitations in segment manipulation, Japanese participants should make few errors when discriminating between ebuço and eboo. In fact, their performance should be indistinguishable from that of French participants. If, in contrast, the perceptual system inserts an epenthetic vowel to break up consonant clusters, Japanese participants should have trouble distinguishing stimuli such as ebuço from stimuli such as eboo because they will in fact "hear" the same thing twice. However, eboo may be "heard" as containing a vowel with different acoustic and phonetic characteristics from the [u] in ebuço. For this reason, in this experiment we chose to have different talkers produce the X stimuli and the other two stimuli (A and B), thereby forcing them to rely on a more abstract and phonological representation rather than on an acoustic and phonetic one. In Experiment 4 we specifically tested the effect of talker change.

Note, however, that comparing the mean performances of different groups of participants (i.e., testing whether Japanese participants were significantly better or worse than French participants on a given task) raises a methodological problem: It is difficult to match populations of participants in all possible respects other than native language. This is why we introduced a complete crossover design in which we made the opposite predictions across the two language groups.

This design was achieved by considering another property of the phonology of Japanese: In Japanese, vowel length is contrastive, such as tokei ("watch") versus tookei ("statistics"). The long vowel is in fact perceived as two adjacent vowels. Therefore, Japanese participants should have no problem performing the ABX task on an ebuço–ebuço contrast. In our stimuli, the ebuço–ebuço contrast had a
range as the ebuso-ebuso contrast (89 ms).

By contrast, in French, vowel length is not contrastive. That is, no pairs of French words can be distinguished purely on the basis of the length of one vowel. The hypothesis under examination was that listeners would impose the phonology of their native language on unfamiliar linguistic stimuli, regardless of whether the stimuli were native or foreign. Hence, we predicted that French participants might have trouble making the ebuso-ebuso contrast but that the Japanese would have no problem.

Method

Participants. Ten Japanese and 10 French participants participated in the experiment. All were recruited in Paris. The age of the Japanese participants ranged from 20 to 48 years (Mdn = 36). Two had no knowledge of French and knew some English. All had been the study of foreign languages before. There were 4 men and 6 women in the group. The age of the French participants ranged from 20 to 50 years (Mdn = 24). None spoke Japanese, but they all had studied English in school. Like the Japanese participants, the French participants had begun studying a foreign language after 12 years of age. There were 9 men and 1 woman in the French group. The Japanese and French participants were all right-handed; they volunteered for the experiment, and no one was paid for his or her participation.

Materials. Sixteen triplicates of the form (ebuso, ebuso, ebuso) were constructed (see the Appendix). All triplicates conformed to the model: V₁C₁CV₂ = V₁C₂CV₂ = V₁C₁CV₂. The first consonants were from the set [b, k, g, j], the final and vowels were from the set [a, i, o, A], and the second consonants were from the set [s, t, d, g, n, m, j, t]. All stimuli were nonwords in both French and Japanese. All stimuli consisted of phonemically valid French syllables and, except for the first member of the triplicates, valid Japanese syllables. Four additional triplicates with the same phoneme range constraints as for V₁C₁, C₂, and V₂ were used in the training set.

The materials consisting of the 20 triplicates were recorded twice: once by a male Japanese speaker and once by a female Japanese speaker. The recordings were made in a sound-attenuated room and digitized at 16 kHz/16 bits on an OROS AU22 digital-to-analog board. Each stimulus was stored in a separate file using a waveform editor. It appeared that although our two Japanese speakers were fluent in French and had some training in phonetics, they could not be prevented from inserting a short vowel [u] within the consonant clusters in some of the ebuso stimuli. These ebuso stimuli were therefore edited with a waveform editor, and the vocalic parts were progressively removed until a French listener found that he or she could no longer hear the [u] vowel. The three classes of stimuli had mean durations of 409 ms for ebuso, 498 ms for ebuso, and 593 ms for ebuso, respectively.

One hundred twenty-eight experimental trials were constructed using the 16 experimental triplicates. Each experimental trial consisted of three stimuli: A, B, and X, where the first two were spoken in a female voice and the last one in a male voice. A and B were taken from the same triplicate but differed in the intermediate vowel duration. There was an ephenthesis contrast (ebuso-ebuso) and a vowel length contrast (ebuso-ebuso). Each contrast could appear in two different possible orders, resulting in four A-B combinations for each triplicate. The X stimulus was identical to either A or B. The overall design was 2 × 2 × 2 (Contrast × Order × Identity). With partial counterbalancing, we obtained 16 training trials using the

Experimental trials were conducted in two blocks, with each condition and item equally represented in each block. The 128 experimental trials were split into two blocks, with each condition and item equally represented in each block.

Procedure. Each experimental trial consisted of the presentation of the three stimuli (A, B, and X), with an interstimulus interval (ISI) of 700 ms. Participants were asked to make a forced-choice judgment of whether X was the same as A or B. Their task was to press a button on their left or right to indicate whether X was the same as A or B. Participants were given 4 seconds to respond. The trial ended immediately after the response or after 4 s had elapsed. The next trial started 1 s later.

In the 16 training trials, participants received feedback as to whether their response was correct. Feedback consisted of the string "correct" or "incorrect" or the string "The response is A" (or B) viewed on a screen. Participants failed to respond before the deadline. Feedback was displayed for 1 s and then was erased from the screen. For incorrect responses, the same trial was repeated immediately until the response was correct. In the two experimental blocks of 64 trials, no feedback was presented. The block number was randomly determined for each individual participant. A short pause was given between the two experimental blocks. Responses were recorded and reaction times (RTs) measured from the onset of the stimulus with the EXPE software package (Pallier et al., 1997).

Results

Four ANOVAs were performed on the results: two per participant and by item and two on error rate by participant and by item (RTs were analyzed only for the correct responses). The ANOVAs had a 2 × 2 design: Language (French or Japanese) × Contrast (ephenthesis, vowel length contrast). The means, standard errors, and error rates are shown for each condition in Table 1.

The analysis of the RT data showed a highly significant interaction between language and contrast, MinF²(1, 29) = 14.16, p < .001. This interaction was due to the fact that the French participants, the vowel length contrast was longer RTs than the ephenthesis contrast (Response mean = 171 ms), MinF²(1, 19) = 12.01, p < .002. For the Japanese participants, there was a trend in the same direction (RT difference = −105 ms), F(1, 9) = 3.77, p = .08; P²(1, 15) = 7.84, p < .02. There was no main effect of language, F(1, 18) < 1, p > .1; F²(1, 15) = 3.81, p = .09. A significant main effect was found for the two-way interaction between Contrast and Language, MinF²(1, 29) = 5.17, p = .03. The effect of Contrast on RT was significant in both languages, F(1, 18) = 22.86, p < .001; P²(1, 15) = 14.36, p < .001.

Table 1

<table>
<thead>
<tr>
<th>Language</th>
<th>RT (ms)</th>
<th>SE</th>
<th>Error (%)</th>
<th>RT (ms)</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>1,082</td>
<td>45</td>
<td>7.5</td>
<td>1,187</td>
<td>75</td>
</tr>
<tr>
<td>French</td>
<td>1,173</td>
<td>73</td>
<td>21</td>
<td>1,002</td>
<td>54</td>
</tr>
</tbody>
</table>

Note. RT = reaction time.

Table 1 Mean Reaction Time, Standard Error, and Error Rate in ABX Judgments on an Epenthesis Contrast and a Vowel Length Contrast in French and Japanese Participants in Experiment 3
$p < .1$, and no main effect of contrast, $F_{(1, 18)} = 1, p > .1$; $F_{(2, 15)} = 5.33, p < .05 < p < .1$.

The analysis of the error data showed the same pattern of results. There was a highly significant interaction between language and contrast, $MinF'_{(1, 26)} = 56.27, p < .001$. This interaction was due to the fact that for French participants, the vowel length contrast was more difficult than the epenthesis contrast, $MinF'_{(1, 13)} = 53.48, p < .001$, whereas the length contrast was easier for the Japanese, $MinF'_{(1, 13)} = 33.48, p < .001$. Overall, Japanese participants tended to make more errors than the French participants, although this was significant only in the items analysis, $F_{(1, 18)} = 3.71, p = .07$; $F_{(2, 15)} = 20.17, p < .001$. Similarly, the epenthesis contrast tended to provoke more errors than the vowel length contrast, but, again, this was significant only in the items analysis, $F_{(1, 18)} = 4.10, p = .058$; $F_{(2, 15)} = 13.18, p < .002$.

**Discussion**

In this experiment, French and Japanese participants had to perform an ABX discrimination task on two contrasts: an epenthesis contrast (ebuzo–ebuzo) and a vowel length contrast (ebuzo–ebuzo). We found a crossover interaction: The Japanese participants had more difficulty with the epenthesis contrast, whereas the French had more difficulty with the vowel length contrast.

These results demonstrate that the phonotactics of a language influence speech perception, even with naturally produced speech stimuli. That is, not only do Japanese participants tend to report more vowels than are really present in the signal (Experiments 1 and 2), but their ability to produce vowels, that is, no matter how familiar the stimuli are, is also strongly affected.

Next, we introduced a change in talker and the two preceding A and B to disrupt participants’ ability to produce consonant clusters and rely on a more abstract representation. However, most researchers have found that non-phonotactic contrasts have no effect on ABX paradigm with no such results still hold without a talker change in which participants can use phonotactics. In the final experiment we investigated the role of talker.

**Experiment 4**

We designed to evaluate the effect of a change in talker on the robustness of the language-specific results. Here, we replicated Experiment 3 and added a new set of stimuli. In this condition, one of the talkers was acoustically identical to the X talker but acoustically different from the A and B talkers. This meant that the X talker would be used at a different time because, in principle, it is the talker that needs to be changed on a purely acoustic basis. If this is still present in the same talker, one participant would not see the X talker and would continue to produce consonant clusters and rely on a more abstract representation. However, most researchers have found that non-phonotactic contrasts have no effect on ABX paradigm with no such results still hold without a talker change in which participants can use phonotactics. In the final experiment we investigated the role of talker.

**Method**

**Participants.** Twenty Japanese participants were recruited (10 in Paris, 8 in New York, and 2 in Nagoya) and tested individually in a quiet room. None of them had participated in the previous experiments. Their ages ranged from 22 to 40 years ($Mdn = 29$). There were 14 women and 6 men in the group.

Twenty French participants recruited in Paris were tested on the same materials. None of them had participated in the previous experiments. Their ages ranged from 19 to 50 years ($Mdn = 21.5$). There were 4 women and 16 men in the group.

Participants filled out a detailed biographical questionnaire about their experience with foreign languages. They also rated their own ability and pronunciation in these languages on a 10-point scale. The questionnaire for the Japanese participants was in English. Moreover, their ability to speak either French or English (or both) was subjectively assessed by a native speaker of French or English, respectively. Four Japanese participants did not fill out the questionnaire.

**Materials.** The same materials as in Experiment 3 were used. We used the same 128 ABX experimental trials of Experiment 3 (A and B stimuli spoken by the male talker) and created another 128 trials with the Stimuli A, B, and X all spoken by the same male talker. In these last trials, X was acoustically identical to either A or B. The overall design was $2 \times 2 \times 2 \times 2$ (Contrast $\times$ Order $\times$ Identity $\times$ Talker). The 256 experimental trials were split into four blocks of 64 trials, with each condition and each item being equally represented in each block.

**Procedure.** The same procedure as in Experiment 3 was used.
Results

The means, standard error, and error rates are shown in Table 2 for each condition. As in Experiment 3, we ran four ANOVAs, two by participants and two by items, on RTs and error rates, respectively, with language, talker, and contrast as experimental factors.

The analysis of the RT data showed that there was a highly significant interaction between language and contrast, $F(1, 53) = 14.81, p < .001$. This interaction was due to the fact that for French participants, the vowel length contrast yielded significantly faster RTs than the epenthesis contrast (117 ms), $F(1, 34) = 14.33, p < .001$, whereas for Japanese participants, there was a nonsignificant trend in the other direction ($-27$ ms, $ps > .1$). No other interaction was significant, except the interaction between language and talker, which was significant only in the items analysis, $F1 < 1; F2(1, 15) = 16.10, p < .001$.

There was a main effect of talker, with the same talker yielding faster RTs than the different talker (85 ms), $F(1, 52) = 12.00, p < .001$. There was also a main effect of contrast, with the vowel length contrast on average yielding slower RTs than the epenthesis contrast (45 ms), $F1(1, 38) = 7.83, p < .01; F2(1, 15) = 5.20, p < .04$. Finally, Japanese talkers tended to have longer RTs than French participants, but this was significantly only in the items analysis (55 ms), $F1 < 1; F2(1, 15) = 18.19, p < .001$.

The analysis of the error data showed similar results. There was a highly significant interaction between language and contrast, $F(1, 40) = 34.11, p < .001$. This interaction was due to the fact that for Japanese participants, the epenthesis contrast yielded significantly more errors than the vowel length contrast, $F(1, 31) = 22.05, p < .001$, whereas for French participants, there was a significant effect in the other direction, $F(1, 33) = 8.62, p < .006$. No other interaction reached significance.

There was a main effect of talker, with the different-talker condition yielding more errors than the same-talker condition, $F(1, 36) = 6.50, p < .02$. There was also a main effect of contrast, with the epenthesis contrast on average yielding more errors than the vowel length contrast, $F(1, 35) = 4.80, p < .04$. Finally, Japanese talkers tended to make more errors than French participants, but this was significant only in the items analysis, $F1(1, 38) = 3.41, .05 < p < .1, F2(1, 15) = 9.51, p < .01$.

Influence of practice. We began our investigation of the effect of practice using a correlational analysis. For each participant, the sequence of RTs on experimental trials was partitioned into 16 successive bins of 16 data points. We found a significant negative correlation between sequential position and mean RT ($R^2 = .67$), $F(1, 14) = 28.10, p < .001$. We also found a significant negative correlation between sequential position and error rate ($R^2 = .67$), $F(1, 14) = 28.02, p < .001$. These effects showed that practice did have an impact and that participants improved their performance with time. We then computed the numerical size of the interaction between language and contrast (i.e., epenthesis-in-Japanese + vowel-length-in-French − epenthesis-in-French − vowel-length-in-Japanese) for each sequential position. There was no significant correlation between sequential position and interaction size either in the RT ($R^2 = .16$), $F(1, 14) = 2.60, p > .1$, or in the error analysis ($R^2 = .17$), $F(1, 14) = 2.91, p > .1$.

In a second step, we ran ANOVAs similar to the ones reported earlier but restricted the analyses to the final block of 64 trials (after 202 trials). In this analysis, the interaction between language and contrast was still significant, both for the RTs, $F(1, 51) = 4.65, p < .04$, and the error data, $F(1, 36) = 17.40, p < .001$.

Influence of language background. Inspection of our questionnaire revealed that the Japanese participants mostly had experience with French or English (one reported having studied some Italian and one some Russian). They all had been taught English as a second language, at an average age of 13.5 years. We divided the participants into two groups, one labeled low proficiency (7 participants) and the other high proficiency (9 participants) on the basis of the means of both their self-evaluation and our evaluation of their pronunciation. High-proficiency participants could understand spoken English or French and sustain a conversation in these languages with good fluency and a modern accent, as assessed by the experimenters. Low-proficiency participants had trouble both understanding and being understood in English or French; some of them did not express themselves in either language.

We found that the proficiency factor introduced a significant effect on the interaction in the RT analysis ($p > .1$). In fact, the high-proficiency group

\[\text{Length Contrast in Japanese Participants and French Participants in Experiment 4} \]

<table>
<thead>
<tr>
<th>Participants</th>
<th>Vowel length (ebugo-ebugo)</th>
<th>Epenthesis (ebugo-ebug)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>SE</td>
</tr>
<tr>
<td>Japanese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same talker</td>
<td>1.008</td>
<td>41</td>
</tr>
<tr>
<td>Different talker</td>
<td>1.058</td>
<td>45</td>
</tr>
<tr>
<td>M</td>
<td>1.033</td>
<td>30</td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same talker</td>
<td>1.095</td>
<td>76</td>
</tr>
<tr>
<td>Different talker</td>
<td>1.225</td>
<td>72</td>
</tr>
<tr>
<td>M</td>
<td>1.160</td>
<td>53</td>
</tr>
</tbody>
</table>

Note. RT = reaction time.

\[4\footnote{The effects on the last block were similar in the same-talker and different-talker conditions for both the reaction time and error rates, although there was a nonsignificant trend toward a larger magnitude of the effect for the same-talker condition (3% reaction times and 1% on the error data).]
roughly the same pattern of errors as did the low-proficiency group (both showed 16% of errors in the epenthesis contrast). In another analysis, we selected the 4 Japanese participants with the greatest proficiency in English or French (both self-rated and as evaluated by an external judge). The selected participants had lived in France or the United States for more than 4 years (1 was an English teacher, 1 was a student of phonetics, and 2 were university students in the United States) and were fluent in French or English. For these participants, the percentage of error on the epenthesis contrast was in the same range as that of the other Japanese listeners (15.9% on average vs. 4.7% for the vowel length contrast).

We also analyzed the linguistic background of the French participants. They all knew English (all had learned it after the age of 6 years), and some also knew German, Italian, Spanish, or Arabic. Note that none of these languages uses vowel length contrastively. However, English, Spanish, and Italian use stress contrastively, and vowel length is used as a cue for stress. We then tentatively divided these participants into two groups (high and low proficiency) according to their evaluation of their proficiency in these languages. We found no effect of fluency on the error data or on the RTs ($p > .1$).

More generally, every Japanese participant we tested in this experiment showed the epenthesis effect (i.e., each participant showed more errors on the epenthesis contrast than on the vowel length contrast). Such regularity was also true of the Japanese participants tested in Experiment 3. In contrast, 18 of 20 French participants (9 of 10 in Experiment 3) showed either no difference or the opposite pattern of behavior. In other words, the observed crossover interaction in the tasks of the different-talker conditions is only found in the same-talker condition.

This apparent discrepancy may be due to the fact that two weak variables can nonetheless conjointly have a significant effect. Indeed, if we look at the first experimental block in the present experiment, we find that the epenthesis contrast yielded 28% errors for the different-talker condition, which is not significantly different from the 34% score of the equivalent first block in Experiment 3. At the onset of both experiments, comparable effect sizes were thus found for the different-talker conditions. In the next trials, however, divergences appear, as the score stayed at 31% in Block 2 of Experiment 3 but dropped to a value centered around 16% in Experiment 4. Such a drop was not found for the same-talker condition, which yielded an initial score of 12% and remained close to this value throughout Experiment 4. In other words, there was an initial difference between the same- and different-talker conditions ($p < .05$), but after exposure to the different-talker conditions, the difference was only found in the same-talker condition.

The General Discussion

Our experiments show that Japanese listeners, tend to perceive the English-speaking Japanese listeners, tend to perceive the English-speaking vowels within consonant clusters. Participants had difficulty discriminating between clusters and did not include a vowel (bhou) versus blocked the task, the different-talker condition produced the same result as the same-talker condition.

In this experiment, we studied the effect of a talker change on the size of the language-specific effects reported in Experiment 3. We found that even though the same-talker condition elicited a small effect on the previously reported interaction between English and French, we found that Japanese listeners had more difficulty with the English contrast than with the French vowel contrasts. This was even more remarkable because the same-talker condition, a judgment of acoustic identity alone was sufficient to perform the task.

We also found that after more than 200 trials the cross-linguistic effects still held. Although practice had a powerful effect on both RT and error rate, it did not significantly modulate the size of the effects.

Finally, a post hoc analysis on linguistic background revealed no clear effect of fluency in languages allowing consonant clusters, such as English or French. That is, both fluent and nonfluent Japanese speakers showed an epenthe-
changes the pattern of data. Needless to say, no tendency
toward epenthesis was present in our French volunteers.

These results buttress the hypothesis that speech perception
is heavily influenced by phonotactic knowledge. This
complements and extends the work by Massaro and Cohen
(1983). Indeed, not only does phonotactic knowledge
influence the classification of the individual phonemes, but it can
also induce the perception of “illusory” phonemes that have
no acoustic correlates. Moreover, it does so in nondegraded
stimuli. This shows that the way in which the continuous
speech stream is segmented into discrete phonemes is not
universal but that it depends on what the typical pattern of
alternation between consonants and vowels is in the
language in question. In brief, when people perceive nonnative
sounds, not only do they assimilate them into their native
categories, but they also may invent or distort segments so as
to conform to the typical phonotactics of their language.

How could such effects be accounted for? We suggest two
possibilities.

The first possibility would be to amend Best’s (1994)
perceptual assimilation model by stipulating that native
categories are not (or not only) categories of single
phonemes but also categories that span larger chunks of signal.
For example, Mehler, Dupoux, and Segui (1990) have
proposed the syllable acquisition, representation, and access
hypothesis (SARAH), a model based on an array of syllable
detectors. In this model, speech sounds are categorized into
syllable-sized units. The repertoire of syllables includes the
totality of the syllables used in the language. Similar
proposals have been made for triphones (Wicklegren, 1969),
diphones (Klatt, 1979), and semisyllables (Dupoux, 1993;
Fujimira, 1976). In such a view, an account of the epenthesis
effect is straightforward. For illustration, we use syllable-
sized categories. Faced with a foreign language, people’s
perceptual system tries to parse the signal using the available
native syllable categories. However, in Japanese, there are
no syllable categories containing consonant clusters or coda
consonants. A stimulus such as /ebzɔ/ therefore activates
categories for “e” and “zo.” It also activates to a lesser
degree all syllables that start with /b/: “bu,” “ba,” “be,”
“bi,” and “bo.” Why is the “bu” interpretation favored?
One possibility is that in Japanese, the [u] vowel is
frequently shortened or devoiced and shows considerable
allophonic variation (see Beckman, 1982; Kentaro & Hoff-
man, 1984). Hence, the log of syllables marked
illuminates the phonotactic constraints.

A second and much different possibility would be to keep
phonemes as the basic level in the perceptual assimilation
model and to add an extra layer of processing that is allowed
to modify the output of the phoneme detectors. For instance,
Church (1987) proposed a parser that would yield a syllabi-
ﬁed representation based on language-speciﬁc constraints.
Indeed, in a study by Pallier, Sebastian-Gallés, Felguera,
Christophe, and Mehler (1993), evidence was found that
listeners build such a syllabified representation on-line (see
also Pallier & Mehler, 1994). To accommodate our data,
such models would have to stipulate that incorrect or deviant
phonological forms are automatically regularized by the
parsing device. The exact nature of the regularization
routines, however, needs to be further specified. Note that
such a proposal would predict a time course difference, and
perhaps a brain localization difference, between phoneme-
based assimilations and phonotactically based assimilations.

Obviously, more studies are necessary to determine which
of these possibilities is correct. However, our findings
already allow us to pinpoint some shortcomings of models
that represent phonemes or subphonemic elements without
any mention of higher order structures (Marslen-Wilson &
Warren, 1994; McClelland & Elman, 1986; Norris, 1994).
In such models, no direct effect of the phonotactic organization
of the language being used is expected in processing, and
our results are difﬁcult to interpret.

Finally, we emphasize the difﬁculty that French partici-
ants experienced in dealing with duration differences. The
French results are interesting because they suggest that not
only the succession of consonant and vowel but also the
precise timing of these elements is important. In Japanese,
vowel length is contrastive, and words can contain up to four
consecutive identical words (e.g., to, too, tooo, toooo). In
French, by contrast, there are no pairs of words that differ
only in vowel length. That a simple acoustic dimension such as
duration has different functions cross-linguistically was
also found by Takagi and Mann (1994), who studied the
perception of English words and nonwords by Japanese
speakers. In particular, they found that in English CVC
syllables, tense vowels are perceived by Japanese listeners
as long vowels (e.g., [ɡɪp] yields [ɡɪpʊ], whereas lax
vowels yield the perception of a geminate consonant (e.g.,
[ɡɪp] yields [ɡɪpʊ]). Hence, the mapping between the
phonetic and phonological level involves more than a set of
phonetic (or even syllabic) detectors but also relies on
rhythmic properties of adjacent phonemes over a large time
window.
Appendix

Materials for Experiments

Materials Used in Experiment 1
abge-abuge, abno-abuno, agmi-agumi, akmo-akumo, ebza-ebuza, egdo-egudo, ibdo-ibudo, igna-ignu, omi-omu, ogza-oguza

gudo, ek(u)-eku(f)-eku(i), efmono-ejuno-efuno, ibdo-ibudo-ibuno, igna-ignu-ignu, ikma-ikuma-ikuma, ifto-ifu(f)-ifuno, obni-obumi-obumi, ogza-oguza-oguza, okma-okuma-okuma, ofa(ofu)-ofa(ofu)

Materials Used in Experiments 2-4
abge-abuge-abuuge, agmi-agumi-agumi, akmo-akumo-akumo, a(ri)-a(jumi)-a(jumi), ebza-ebuza-ebuza, egdo-egudo-

Received September 16, 1997
Revision received September 1, 1998
Accepted October 20, 1998