Phonological Knowledge and Perceptual Epenthesis*

Korlin Bruhn, Bridget Samuels & Bert Vaux

1 Introduction

It has long been known that one's native (L1) phonemic inventory can influence the perception of non-native (L2) contrasts (e.g., Sapir 1933). While infants initially display the ability to distinguish both native and non-native phonemic contrasts (cf. Saffran et al. 2006), by around 10–12 months (e.g., Werker and Tees 1984a, Werker 1989) this ability is significantly attenuated as the child becomes attuned more specifically to the L1 inventory. More recently, the difficulty of perceiving sequences that are phonotactically illicit in L1 has also been widely discussed: studies have claimed that French-speaking listeners perceptually assimilate illicit [tl], [dl] onset clusters to licit /kl/, /gl/ onsets (Hallé et al. 1998); that English listeners epenthesize a schwa to repair onset clusters that violate the language's preferred rising sonority profile (Berent et al. 2007, Davidson et al. 2007); and that Japanese speakers use epenthesis to break up illicit consonant clusters and resyllabify coda consonants in loanwords (e.g., Dupoux et al. 1999).

In the present work we focus on this phenomenon of epenthesis in Japanese and its relevance for models of speech perception. We present evidence that Japanese speakers can still access phonetic details necessary to learn a non-native contrast, in spite of their native phonology. Japanese disallows all consonants other than nasals and the first part of geminate consonants in coda position. Consequently, foreign loanwords violating these phonotactic restrictions are repaired via epenthesis of /u/ (typically pronounced [u̟ɯ], as will become relevant later) or /o/ (data from Itô and Mester 1995):

(1) a. 'Sphinx' → *sufiŋkusu*
b. 'Zeitgeist' → *tˢaitogaisuto*
c. 'fight' → *faito*

* The present work owes a great deal both directly and indirectly to Mark Hale's thinking on perception, acquisition, the initial state of the language acquisition device, and what performance effects do and do not reveal about competence. The third author remembers vividly discussions with Mark in the basement of Grays Hall and the ground floor of 77 Dunster Street in the 1990s about contrast-driven feature acquisition versus full perceptual access, ideas which were eventually published as Hale and Kissock (1998), Hale et al. (2007) and Hale and Reiss (2008), and (along with Janet Werker's findings and theories, whose significance Mark and Charles Reiss first presented to us) significantly shaped the thinking reflected in the present chapter. We appreciate helpful suggestions from Laura Grestenberger, Charles Reiss, and Markus Pöchtrager, which greatly improved the paper.

Dupoux et al. (1999) draw a connection between perception and loanword adaptation, suggesting that adaptation happens at the perceptual level: That is, when Japanese listeners are confronted with, e.g., [sfɪŋks], they perceive it as [sɯ̥.ɸiŋ.kɯ̥.sɯ̥]. This seems to be confirmed by the results of two of their experimental studies (Dupoux et al. 1999, Dupoux et al. 2001), in which French and Japanese subjects indicated whether they heard a medial /u/ in nonword stimuli taken from a VCCV-VCuCV continuum. Japanese listeners indicated the presence of the vowel around 70% of the time even in the VCCV condition. In two speeded ABX tasks—one with and one without speaker change—Japanese listeners had difficulty distinguishing stimuli like *ebzo* and *ebuzo* reliably. Japanese listeners made fewer errors in the same-speaker task than in the different-speaker task but still made significantly more errors than the French listeners. Dupoux et al. concluded that phonotactic knowledge influences listeners so strongly that it creates a perceptual illusion: the Japanese listeners judge there to be a speech segment (/u/) despite the absence of acoustic correlates in the signal, simply because their L1 phonology insists it should be there.

Subsequent studies by Dupoux and colleagues pursued this so-called *ebzo*-effect. In a 2001 study, they evaluated lexical influences (cf. Ganong 1980): In a speeded lexical decision task, participants judged whether a CVCCV speech token was a real word or not. The stimuli were designed so that insertion of /u/ or a different vowel would yield an existing word: e.g., *sokdo* is not a word of Japanese, but *sokudo* means 'speed'; *mikdo* is not a word but *mikado* means 'emperor.' The stimuli that needed an epenthetic /u/ to form a word were largely judged to be real words despite the surface consonant cluster. The stimuli that required a vowel other than /u/ to form a word were mostly judged to be non-words. The authors thus concluded that the illusion of epenthesis cannot be caused by top-down influences from lexical neighbors, but rather must be a pre-lexical effect; perceptual insertion of the vowel happens before the lexicon is consulted.

Dehaene-Lambertz et al.'s (2000) ERP study using an oddball paradigm with stimuli such as *ebzo* and *ebuzo*, which contrast after their first consonant, appears at first blush to confirm that phonology influences speech perception at very early processing stages. French but not Japanese listeners displayed a mismatch negativity (MMN) at a latency of 140–280 ms after the offset of the oddball's first consonant. As MMNs are usually elicited when a change in the stimulus is noticed, the lack of MMN in the Japanese listeners in response to an *ebzo-ebuzo* change implies that they do not perceive these stimuli to be different from one another. This suggests that epenthesis of the vowel happens at a stage so early in processing that it prevents an MMN from being elicited.

Why does this perceptual epenthesis occur? The answer to this question must be formulated within the larger context of Japanese phonology. While Dupoux et al. (1999) and Dehaene-Lambertz et al. (2000) only tested perceptual epenthesis of /u/, vowels other than /u/ can be found inserted in loanwords: /o/ is typically inserted after dental stops (e.g., 'fight' borrowed as *faito*), presumably because when a /u/ follows a dental stop, the stop becomes affricated, e.g., /kat-/ 'win' → [katanai] 'win - negated' but [kat^{s}ɯ] 'win - present tense' (Itô and Mester 1995). Only /u/ appears to be perceptually inserted, though, as a study by Monahan et al. (2009) suggests: nonword stimuli like *etma*, which offer the environment for an /o/ to be inserted,

could be distinguished from both *etuma* and *etoma* by Japanese listeners, whereas participants had trouble with the classic *ebzo-ebuzo* contrast. This suggests that /u/-insertion is blocked in environments where /u/ is not phonologically licensed, yet nonetheless /o/ is not perceptually inserted. One explanation suggested in the literature (Dupoux et al. 1999, Dupoux et al. 2001, Dupoux et al. 2011) is that high vowel devoicing, which is optional in many dialects (cf. Vance 1987, chapter 6), plays a role. This means that both /i/ and /u/ can be realized as anything from a whispered vowel with visible formants to simply frication without formant information, or even total deletion (Tsuchida 1987, Varden 1998). On the other hand, /o/ does not typically undergo devoicing,[1] so it might be sufficiently distinguishable from [Ø] that its presence cannot be posited without strong acoustic evidence in the signal—cf. also Steriade's (2001, 2009) P-map theory, in which the perceptually minimal deformation of the input is chosen in phonotactic repairs, and Samuels and Vaux (2020) on the silence-cued perception of epenthetic stops. A similar argument could rule out the perception of /u/ in stimuli like *etma*, since /u/ triggers affrication of a preceding dental stop: [t^{s}u] and even [t^{s}u̥] could be too perceptually different from [tØ] to be perceived here.

Another important question is why /u/, but not the other Japanese high vowel, /i/, is subject to perceptual epenthesis.[2] We return to this issue in Section 4.1. For now, we raise the following question: if the only reason for perceptual epenthesis is that these vowels can sometimes be realized as [Ø] and thus that listeners are used to interpreting CC-sequences as underlyingly containing a high vowel, then why is /u/ preferred over /i/? Further, high vowel devoicing usually happens between two voiceless consonants, yet the stimuli used in Dupoux's studies mainly involve environments that do not license devoicing. The key could be that /u/ in Japanese is the closest to [Ø] in that it is the shortest vowel (Beckman 1982) and the one which allows the most formant variability (Keating and Huffman 1984).[3] This leads Dupoux et al. (1999, 2001) to a modified version of Best's (1994) Perceptual Assimilation Model, wherein they suggest that the perceptual unit in Japanese is the syllable. Therefore, when a foreign sequence is perceived, it is mapped syllable-by-syllable onto the perceptually closest native category, which for CØ would be C/u/ for the reasons set out above.

Peperkamp and Dupoux (2003) and Peperkamp (2005) refine this theory into what we will refer to as the Phonetic Decoder model. In this model, a phonetic decoding module maps the speech signal, one word at a time, into a discrete phonetic representation that conforms to the L1 phonology; a phonological decoding module then maps this surface form onto an underlying representation. Accordingly, a sequence that is phonotactically illicit in the L1 cannot be mapped accurately: The phonetic decoder for L1 Japanese speakers cannot accommodate two consonants next to each

1 Nonhigh vowels sometimes devoice, too, but less often and less systematically (Vance 1987: 48f.) and will therefore not be considered here.

2 It should of course be noted that epenthetic—and specifically perceptually epenthetic—vowels differ cross-linguistically. For example, Dupoux et al. (2011) demonstrate that in a context where Japanese listeners indicate the presence of a perceptually epenthetic /u/, Brazilian Portuguese listeners perceive an epenthetic /i/.

3 Alternatively or in addition, this variability could indicate that /u/ is featurally underspecified. We set this issue aside.

other, so an empty vowel segment intervenes and is carried over to the phonological mapping, where the empty vowel slot is interpreted as the closest phonetic match.

However, in Dupoux et al.'s (1999) study an effect of condition (i.e., different responses to *ebzo* and *ebuzo*) was observed in Japanese listeners for an ERP at a latency of 290–400 ms, suggesting that speakers perceive the difference between stimulus types on some level. Moreover, Tremblay et al. (1997) have shown that a weak MMN response to non-native contrasts like the one Dehaene-Lambertz et al. (2000) found for Japanese listeners can be strengthened by training. Tremblay et al. trained native English speakers on the non-native category of pre-voiced labial stops and found electrophysiological evidence that these subjects could generalize the newly learned VOT boundary to pre-voiced alveolar stops.

Ample behavioral evidence also points to the conclusion that some ability to perceive non-native phonological patterns remains into adulthood. Werker and Tees (1984b) found that under certain circumstances, even difficult foreign contrasts can be distinguished by adults. Furthermore, when listeners are not in "speech mode" or when auditory linguistic stimuli are altered to a sufficient degree that renders them non-speech-like, the ability to discriminate non-native contrasts reveals itself as intact (Best et al. 1981, Remez et al. 1981, Liberman 1982, Werker and Tees 1984b). Davidson et al. (2007) found that a picture-matching task which teaches participants to distinguish minimal pairs can help English listeners to overcome epenthetic perceptual repair of illegal onset clusters. By associating each stimulus with a meaning—e.g., *zemagu* denotes a picture of a dragon while *zmagu* refers to a fish—performance was enhanced relative to an AX discrimination task using the same stimuli without associated meanings. Even in Dupoux et al.'s (1999) study, while Japanese listeners underperformed relative to the French listeners, they performed significantly better than chance in the VCCV vs. VCuCV perception task, and their error rate was also lower than would be expected if they were simply guessing in the ABX task. In sum, the evidence suggests that the perception of non-native contrasts persists on some level in adults and can be improved. Thus, perceptual illusions of the *ebzo-ebuzo* type may not be so inevitable after all.

The question therefore arises: Can Japanese listeners be taught to perceive faithfully consonant clusters such as the one in *ebzo*? If they are indeed able to avoid or override perceptual epenthesis, it would show that Japanese listeners are not "'deaf' to the difference between *ebuzo* and *ebzo*" (Dupoux et al. 1999: 1574) and imply that phonological influence is not as inevitable as Dupoux and colleagues assert. If Japanese listeners, like the English listeners studied by Davidson et al. (2007) (cf. also Berent et al. 2007), can be taught not to activate perceptual epenthesis, it would indicate that even in those cases where phonotactic knowledge has been claimed to interfere with speech perception at very early stages (e.g., Dehaene-Lambertz et al. 2000), phonetic detail must still be perceived accurately initially before phonology alters the percept. Otherwise, Japanese listeners would not be able to use the phonetic detail to learn that there is a meaningful difference between *ebzo* and *ebuzo*.

If Japanese speakers are indeed able to learn such non-native contrasts, it would provide some evidence against the Phonetic Decoder model. The present study therefore investigates whether perceptual epenthesis can be overcome by Japanese listeners by adapting Davidson et al.'s (2007) study to the sound patterns in question. We also

replicate Dupoux et al.'s (1999) ABX task. This allows for a more reliable measure of improvement while at the same time allowing a direct comparison to Dupoux et al.'s study. We demonstrate that Japanese speakers are able to learn the *ebzo-ebuzo* contrast, and provide an explanation of these results that does not depend on the Phonetic Decoder model.

1.1 Predictions

Both possible outcomes of this study—the contrast between *ebzo* and *ebuzo* can or cannot be taught to Japanese L1 speakers—have interesting implications and each outcome is predicted by a different class of models of speech perception.

1. Dupoux and colleagues (e.g., Dupoux et al. 1999, 2001, 2011; Dehaene-Lambertz et al. 2000; Peperkamp and Dupoux 2003) predict that Japanese listeners cannot be taught to perceive the *ebzo-ebuzo* contrast reliably. That is, a contrast can only be perceived if L1 phonotactics allow it.[4] If this prediction is correct, then the triggering factors for English onset epenthesis (e.g., *lbif-lebif* in Berent et al. 2007) must be different from those involved in the Japanese case to explain the fact that English listeners can learn to override perceptual epenthesis (Davidson et al. 2007) but Japanese listeners apparently cannot. This is consistent with the Phonetic Decoder model—L1 categories must be available for the speech signal to be mapped onto. The relevant underlying difference between English and Japanese may be represented in terms of CV sequences (cf. also Clements and Keyser 1983). Japanese listeners would thus map a stimulus like *ebzo* onto V.CV.CV, because no VC.CV template is available. English listeners, however, have an appropriate template available to parse the CC sequences in both *ebzo* and *lbif* accurately: while they may not be used to mapping *lb* to an onset sequence, their task under this interpretation is merely to map *lbif* onto the CCVC template available in English, rather than onto the CVCVC template which would create a percept of *lebif*.
2. The alternative hypothesis holds that Japanese listeners can be taught to overcome perceptual epenthesis reliably. This outcome would not be consistent with the Phonetic Decoder model. Instead, it would suggest that, in certain tasks, Japanese listeners can access the phonetic details which provide the crucial information to distinguish *ebzo* from *ebuzo*, especially given that the phonemes involved in the contrast are already in the L1 inventory of Japanese listeners.

1.2 Approach

The current study applies Davidson et al.'s (2007) paradigm to Japanese speakers, testing whether discrimination between the types of stimuli that Dupoux et al. (1999 et seq.) have used in their studies can improve when the crucial phonetic difference is

4 Note that this does not mean that all consonant clusters are treated equally: some are misperceived more often than others. For example, Berent et al. (2007) found that word-initial sonority falls are misperceived more often than sonority plateaus. Accidental gaps in a language's inventory of allowed sequences should be relatively unproblematic for perception, by virtue of being accidental rather than systematic.

highlighted. A brief explanation of the paradigm is necessary in order to understand the motivation behind the changes that we introduced. The methods are reviewed in detail in Section 2.

Davidson et al.'s (2007) experiment is a three-phase picture-masking task (PMT). In the familiarization phase, participants are introduced to picture-name pairs (see Figure 2). In the following training phase, participants receive training on contrasts that are difficult for them to perceive correctly; the training involves matching minimal pair stimuli to the pictures they denote. Participants hear a name and have to match it to the correct picture (Figure 3). Once they have correctly matched all of the names in one trial without a mistake, participants move onto the test phase in which they see a picture and hear the minimal pair together and are asked to indicate which one of the two minimally contrasting stimuli denotes the picture (Figure 4). To ensure that the contrast learned in the training phase for one speaker can also be carried over to another speaker, a different voice is used in the test phase.

The present study adds two ABX tasks to the PMT, for two reasons. One motivation was to monitor the effect of training more reliably, as the original PMT design does not include a measure of performance before training on the contrast has occurred. Davidson et al. (2007) instead refer to a previous AX-discrimination experiment with the same stimuli (Davidson 2007). Participants performed at chance in that task, which compared CC onset stimuli with their counterparts containing actual schwas (e.g., [vtɑke]∼[vətɑke]). Because participants chose the correct token 60.5%[5] of the time in the PMT (Davidson et al. 2007), the authors hypothesized that their learning paradigm was successful.

There are several issues with using a different task from a different experiment as a baseline for performance improvement. Firstly, this comparison interprets a performance difference between two different groups of participants as 'improvement'. Secondly, the AX experiment (Davidson 2007) and the PMT (Davidson et al. 2007) are not directly comparable, because they test different abilities. The AX task required participants to indicate whether two stimuli heard consecutively were "exactly the same" or different, thus placing emphasis on acoustic identity rather than phonological category membership of the two items. However, in order to perform well in the PMT test phase, participants had to build phonological representations of minimal pairs, since they had to carry over the learned contrast to a new speaker. Consequently, the difference in performance between those two tasks could stem from the difference in tasks. The present study eliminates these confounds by running the same ABX task before and after the PMT. The 'before' task provides a baseline against which performance in the 'after' task can be evaluated; an increase in performance is thus likely to be a result of the training.

Given the methodological concerns just discussed, we felt it was necessary to replicate the training paradigm used by Davidson et al. (2007) with English-speaking listeners to confirm its efficacy. This group was tested on word-initial sequences that typically trigger schwa epenthesis for English listeners. Consequently, if the

5 This figure was not explicitly given in Davidson et al. (2007). Their analysis involved dividing the participants into "high performer" and "low performer" groups as well as splitting the scores according to whether the target word had a CC or a CəC onset. When these groups/conditions are collapsed, the overall percentage of correct responses is 60.5%.

English group showed performance improvement, a negative result for Japanese listeners would not indicate a faulty training method but could be attributed to the insurmountable influence of phonology on perception.

The second motivation for the ABX tasks in the present study was to enable comparison to parts of Dupoux et al. (1999). Since the present study's intention is to test whether Japanese listeners can learn to suppress perceptual epenthesis and thus improve performance on the *ebuzo/ebzo*-type contrast, it is advisable to use the more challenging task as a baseline in order to maximize possibility for improvement. Of the two ABX designs that Dupoux et al. (1999) used, the more challenging one (with an error rate of 32%) was Experiment 3, which used one speaker for A and B and a second speaker for X. The speaker change is designed to prevent participants from using only acoustic cues; they have to build a more abstract, phonological representation of the stimuli, so that X can be compared to A and B despite the speaker difference. Given that it has been argued especially for the Japanese case (e.g., Dupoux et al. 1999, 2001; Dehaene-Lambertz et al. 2000) that phonology interferes with speech perception at even very low levels of acoustic processing, it can be assumed that a task in which phonological representations are built and compared allows even more phonological interference to distort the acoustic input. If, as predicted by the Phonetic Decoder model, both stimuli A and B map onto the same phonological form (i.e., VCuCV), participants should have trouble matching X with the correct stimulus. On the other hand, if Davidson et al.'s (2007) PMT training paradigm is successful, then Japanese participants should be able to learn to suppress the epenthesis percept.

In addition to the fundamental question of whether Japanese listeners can overcome perceptual vowel epenthesis, the present study also addresses the basis for vowel epenthesis. Dupoux and colleagues have frequently evoked the notion of high vowel devoicing as a possible factor in perceptual vowel epenthesis (Dupoux et al. 1999, Dupoux et al. 2001, Dupoux et al. 2011) but did not examine this possibility systematically, instead only analyzing their results retrospectively. Consequently, their stimuli were not balanced for contexts that are known to be unfavorable to high vowel devoicing, which is unfortunate because this bears on whether the Phonetic Decoder model adequately describes the observed results. If favorable contexts yield more epenthesis, this could be taken as confirmation that acoustic proximity (cf. Peperkamp and Dupoux 2003) does indeed play a role in the mapping of the acoustic stream onto native syllables. For example, with a stimulus like *ebzo*, we would not expect much perceptual epenthesis because high vowels do not devoice between voiced consonants. On other hand, with a stimulus that presents a more favorable context for /u/-epenthesis, such as *etma*, we would expect at least some Japanese listeners to accept CØ as a variant of C/u/ such that *etma* is not as perceptually distinct from *etuma* as *ebzo* is from *ebuzo*.

To investigate this, 50% of the present study's stimuli contain a phonetic context favorable for high vowel devoicing, i.e., both surrounding Cs are voiceless obstruents, and the other half contains contexts which are unfavorable to high vowel devoicing, i.e., at least one of the Cs is voiced. If high vowel devoicing plays a role in vowel epenthesis, we predict that the phonetic contexts triggering devoicing should elicit significantly more perceptual epenthesis than the other environments, and hence that

Japanese listeners will make more mistakes distinguishing the consonant cluster from its epenthetic counterpart.

2 Materials and Methods

2.1 Participants

2.1.1 English

Fifteen monolingual native speakers of British English (3 men and 12 women) participated in all three tasks (ABX1, PMT, ABX2). None reported any hearing impairments. The average age was 22.8 years (SD = 2.27). Two participants had received brief training in phonetics while attending university but their performance did not deviate from that of participants without phonetic training; their scores were therefore included. Four participants had experience with languages allowing the onset clusters used in this experiment, namely Hebrew and Russian, but began study relatively late (at ages 16, 18 and 22).[6] One of these participants also had received Welsh lessons in early childhood but claimed that it had not been extensive enough to learn the language; in any case, Welsh does not allow the onsets used in this experiment. None of these participants were outliers with regard to their performance, suggesting that experience with those languages was not an influencing factor.

2.1.2 Japanese

Ten Japanese native speakers (5 men and 5 women) raised monolingually in Japan participated in all three tasks (ABX1, PMT, ABX2). None reported any hearing impairments. The mean age was 25.6 (SD = 4.84). Except for one participant, who started learning English at the age of 11, none had learned a foreign language extensively before the age of 12. One other participant began studying French at the age of 8 but claimed that it had not been extensive enough to learn the language properly. One participant reported being diagnosed with Asperger's Syndrome but both his results and the time he needed to complete the experiment were within the range of the other participants, so we decided not to exclude him. The average time that participants had lived in an English-speaking country was 4.1 years (SD = 3.59, range = 10.5, mode = 2).

2.2 Materials

All recordings were made in a sound-attenuated booth onto a Nagra Ares-M II handheld digital recorder with a cardioid Sennheiser ME64 microphone at a sampling rate

6 Pallier et al. (1997) found that even adult Spanish-Catalan bilinguals who had been exposed to L2 on a daily basis before the age of six did not show categorical perception on the Catalan phoneme contrast /e/-/ɛ/. While this study was concerned with phonemes rather than phonotactics, it nevertheless suggests that even language acquisition in early childhood does not necessarily result in complete native-like language competence. Therefore, language learning during or after adolescence should not influence the results. Note also that our participant pool is similar to that of Dupoux et al. (1999), which included participants who had begun learning a foreign language after the age of 12.

of 22050 Hz. Recordings were edited as described below in Praat, and any recording artifacts (clicks) were excised from the sound files. Every stimulus was assigned to a picture of a unique cartoon character (16 experimental pictures and 6 practice pictures), so that participants learned to associate the auditory stimuli with meanings.

2.2.1 Practice set

Three German minimal pairs differing in the initial consonant were recorded once by a female native German speaker and once by a male native British English speaker who is also fluent in German: *Scholle-Wolle, Sonne-Tonne, Mund-Hund*. These stimuli were used in the practice sessions of all three tasks to familiarize participants with the task. This is the same approach used by Davidson et al. (2007) in their PMT, using Spanish words rather than German ones.

2.2.2 English set

For the English stimulus set, the same nonce words, recorded anew for this experiment, were used in all three tasks and were a subset of those used by Berent et al. (2007). Since the Japanese experimental design included investigation of two different conditions (contexts favorable and unfavorable to devoicing high vowels), the English experimental design also included two different conditions of differing favorability to epenthesis, namely one with onsets with sonority plateaus and a second with falling sonority, as the latter trigger schwa epenthesis more often than the former (Berent et al. 2007).

A set of four tokens with sonority plateau (= SP) onsets (e.g., *kpim*) plus their epenthetic counterparts (e.g., *kəpim*) and four with sonority falls (= SF) (e.g., *mkag*) as well as their counterparts with inserted schwas (e.g., *məkag*) were used. A male native British English speaker recorded the eight CəC-onset tokens and the eight corresponding CC-onset tokens were created by excising the schwa from the recorded CəC set. In addition, the entire set of nonce words was recorded by a female native British English speaker as detailed below. A native British English speaker and trained phonetician listened to all of the stimuli (male and female) and deemed them natural-sounding.

To ensure that listeners were distinguishing the stimuli within a pair on the basis of the contrast under investigation (presence versus absence of schwa), CC-onset stimuli were made from recordings of CəC-initial stimuli, to generate pairs of stimuli that were acoustically identical apart from the schwa. This set was used to train participants on the contrast: A and B in the ABX tasks were drawn from it, and it was used in the training phase of the PMT. CC onset stimuli were created using Praat (Boersma and Weenink, 2011) to digitally remove the schwas from the CəC recordings at zero crossings. Care was taken not to create abrupt transitions that could result in unnatural-sounding stimuli. Since the length of the schwa plays an important role in the discriminability between CC and CəC, the CəC stimuli were edited to have schwas of average length, as measured by Davidson (2007) and employed by Davidson et al. (2007), namely 68 ms. Some of the schwas in the CəC stimuli thus had to be reduced in length slightly, by cutting single periods at zero crossings. If more than

one period had to be removed, non-neighboring periods were selected, so as not to obscure formant transitions.

Stimuli recorded by the female speaker were used for the X in the ABX tasks and in the test phase of the PMT. Acoustic identity was not an issue for this set, since it was not used in the training phase. Therefore, naturalness of the stimuli was favored over consistency with the male set. Further, we wished to avoid the possibility of participants simply exploiting the length difference of the stimuli if both sets of stimuli used near-identical CC-CəC pairs. By using different recordings for the female speaker, we sought to ensure that participants would have to learn the abstract contrast in order to perform well. Thus, CC stimuli were recorded naturally. Schwas were reduced or lengthened to around 68 ms (M = 63 ms) by removing or inserting single periods at zero crossings. In some cases the onset Cs of the CəC recordings had to be spliced and used as the onset C of the CC-stimuli because they differed strongly in quality. For example, the speaker produced very strong aspiration in /tpʌg/, so the [t] from [təpʌg] was used to replace the original one, and since the [m] in [mkʌg] was longer than the entire [mə] in [məkʌg], a few periods were excised from it.

2.2.3 Japanese set

For the Japanese stimulus set (recorded anew for this study), the same tokens were used in all three tasks and were a subset of the stimuli used by Dupoux et al. (1999) except for the nonwords *aksa-akusa*, which were added since otherwise an insufficient number of stimuli included phonetic environments that trigger high-vowel devoicing. Four tokens with phonetic contexts favorable to high-vowel devoicing (= FC) and their epenthetic counterparts (e.g., *ekshi-ekushi*) and four tokens with unfavorable contexts (= UC) and their epenthetic partners (e.g., *ebza-ebuza*) were recorded once by a male native British English speaker and once by a female native British English speaker.

While making the Japanese stimuli, the main concern was to eliminate the possibility of coarticulatory cues betraying the former presence of a vowel after digitally removing the [u] from a VCuCV token. (Note that Dupoux et al. 1999: 1573 used Japanese speakers who "could not be prevented from inserting a short vowel [u] [. . .] in some of the *ebzo* stimuli.") Dupoux et al. (2011) showed that Japanese listeners are highly sensitive to such coarticulatory traces, and it has further been shown that even 10 ms of a stop burst can contain enough information to identify a following vowel (Blumstein and Stevens 1980). In order to avoid these problems, both VCuCV and VCCV tokens were recorded naturally.

To stay as close as possible to Dupoux et al.'s (1999) study, the average vowel length of that study—89 ms[7]—was used, and vowels were accordingly reduced by removing single periods at zero crossings or lengthened by inserting single periods: the average length was 89 ms for the male set and 90 ms for the female set.

7 Dupoux et al. (1999) do not explicitly state the length of the /u/s in their stimuli. They only indicate that the mean difference between *ebzo* and *ebuzo* items was 89 ms, suggesting that the difference is due to the presence vs. absence of the vowel. However, it is entirely possible that other elements in the stimuli varied in length as well. Nevertheless, since this is the only indication the authors provide regarding the crucial length of the vowel, we decided to use this as target length for the /u/s in our study.