# Effects of language experience on pre-categorical perception: Distinguishing general from specialized processes in speech perception

Paul Iverson,[a] Anita Wagner, and Stuart Rosen
*Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom*

(Received 13 August 2014; revised 23 February 2016; accepted 11 March 2016; published online 13 April 2016)

Cross-language differences in speech perception have traditionally been linked to phonological categories, but it has become increasingly clear that language experience has effects beginning at early stages of perception, which blurs the accepted distinctions between general and speech-specific processing. The present experiments explored this distinction by playing stimuli to English and Japanese speakers that manipulated the acoustic form of English /r/ and /l/, in order to determine how acoustically natural and phonologically identifiable a stimulus must be for cross-language discrimination differences to emerge. Discrimination differences were found for stimuli that did not sound subjectively like speech or /r/ and /l/, but overall they were strongly linked to phonological categorization. The results thus support the view that phonological categories are an important source of cross-language differences, but also show that these differences can extend to stimuli that do not clearly sound like speech. © *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4944755]

[TCB] Pages: 1799–1809

## I. INTRODUCTION

Speech perception develops such that individuals become specialized to discern differences between native-language (L1) phonemes, and this specialization likely interferes with the learning of new second-language (L2) phonemes as an adult (e.g., Best, 1995; Flege, 1995; Kuhl and Iverson, 1995). Researchers have traditionally put the locus of this L1-L2 interference at the level of phonetic or phonological categories. For example, theorists going back to the 1930s have suggested that individuals perceive L2 speech sounds through the filter of their L1 phonology (e.g., Trubetzkoy, 1969). Best's (1995) Perceptual Assimilation Model states that listeners perceive non-native speech sounds in terms of how they assimilate to L1 phonological categories. Flege's (1995) Speech Learning Model states that L1 and L2 phonetic categories exist together rather in separate subsystems of phonetic processing, making it difficult to learn a new L2 category when it is too close to an existing L1 category.

However, it is not clear that phonetic or phonological categorization is necessary during speech recognition. There were early claims that categorical phonetic encoding is critical to the speech recognition process (e.g., Studdert-Kennedy *et al.*, 1970), but even at the time it was known that phonetic encoding could not account for all of speech perception (see Repp, 1984 for a review). More recently, it has been found that phoneme labeling cannot explain sensitivity to dimensions orthogonal to categorization boundaries or among stimuli labeled the same (e.g., Iverson *et al.*, 2003, 2008), the link between identification and discrimination is task dependent (e.g., Schouten *et al.*, 2003), and within-category variation appears to be represented continuously relatively late into neural processing (e.g., Toscano *et al.*, 2010). Word recognition research suggests that form-based lexical representations are phonetically detailed rather than being based on phoneme labels (e.g., Andruski *et al.*, 1994; McMurray *et al.*, 2002), and that individuals probably do not need to categorize phonemes on the way to recognizing words (e.g., Norris *et al.*, 2000; Norris and McQueen, 2008). Likewise, many studies have demonstrated that listeners are able to remember and adapt their speech recognition processes to individual differences among talkers, in ways that make it clear that listeners are sensitive to finer-grained differences than category labels (e.g., Nygaard and Pisoni, 1998).

Given this uncertain role of categorization in speech perception, it seems unlikely that it can explain the majority of L2 perceptual difficulties, unless the notion of categorization is broadened to include a range of processes between sensation and contact with the lexicon, rather than only being a labeling/decision/encoding stage. However, the possibility that language experience can affect relatively early processes is conceptually difficult, because it can blur the traditional divisions between general perceptual and speech-specific processes (e.g., Diehl and Kluender, 1989; Liberman and Mattingly, 1989; Remez *et al.*, 2001; Stevens and Blumstein, 1978; Werker and Curtain, 2005; Whalen, 1997). For example, several studies by Krishnan *et al.* (2005, 2009a, 2009b) have suggested that the frequency following response generated in the auditory brainstem more accurately tracks the contour of Mandarin tones for Mandarin speakers than for English speakers, and that this difference

[a]Electronic mail: p.iverson@ucl.ac.uk

does not occur for language-neutral variation in pitch. If language-experience effects occur this early, then it becomes doubtful that there exists a purely general level of perception, at least beyond the cochlea, which feeds into speech-specific processes. On the other hand, pitch-tracking accuracy in the brainstem seems conceptually more like an aspect of sensory processing than anything that would usually be called categorization or attributed to a speech-specific processing module (e.g., Liberman and Mattingly, 1989).

The present study explored this territory between general auditory and speech-specific processing by comparing the perceptual sensitivity of adult English and Japanese speakers for /r/-/l/ stimuli with varying acoustic forms. Adult native Japanese speakers have difficulty learning to produce and perceive these phonemes when learning English, and are much worse than English speakers at discriminating /r/ and /l/ stimuli that cross the English identification boundary (Goto, 1971; Miyawaki *et al.*, 1975). However, Miyawaki *et al.* (1975) found that this cross-language discrimination difference was eliminated when listeners heard only isolated F3 stimuli (i.e., stimuli which consist solely of the critical acoustic difference between /r/ and /l/, but which sound like chirps rather than speech). Their result contributed to the view that the phonological categorization of these stimuli produced cross-language differences, and that the basic auditory processing of F3 was the same. Although this seemed like a clear conclusion, subsequent work has demonstrated that Japanese listeners specifically have difficulty with F3 in /r/ and /l/ rather than having more general problems with acoustic dimensions that affect the assimilation of these phonemes into L1 Japanese categories (e.g., transition duration), and that individual differences in L1 category assimilation do not predict L2 English categorization (Hattori and Iverson, 2009). It may be premature to rule out cross-language auditory processing differences based on the data of Miyawaki *et al.* (1975), because there is a large acoustic difference between a full /r/-/l/ syllable and an isolated F3 transition, leaving a relatively unexplored range of stimuli that vary in terms of their acoustic similarity, and subjective phonetic similarity, to natural speech (e.g., Rosen and Iverson, 2007). It may be that cross-language differences in F3 sensitivity occur when an F3 transition is embedded within complex stimuli that resemble /r/ and /l/ syllables but fall short of sounding like intelligible speech.

This approach of exploring the territory between speech and non-speech was used by Iverson *et al.* (2011) to examine the perception of English /w/-/v/ by English and Hindi speakers. We used a set of English /w/-/v/ continua that acoustically deconstructed the stimuli so that they varied in the degree that they sounded like speech or non-speech. The stimuli manipulated the carrier spectrum (normal speech with natural pitch contour, normal speech w/ flat pitch, sawtooth spectrum), the presence/absence of formant movement, normal or flattened amplitude contour, and the presence/absence of frication amplitude contrast, such that the stimuli sounded like natural /w/ and /v/ with the most natural combination of these acoustic manipulations, but became progressively less categorizable and less subjectively like speech for less natural combinations. We found cross-language differences for stimuli that sounded clearly like /w/ and /v/ (i.e., English speakers had higher discrimination sensitivity) and no cross-language differences for stimuli that sounded unlike speech. However, there were also cross-language differences for a single set of stimuli in between, that sounded mostly like speech but whose phonemes could not be reliably identified. We thus concluded that cross-language differences may occur only when the stimuli are perceived in a speech mode (e.g., Mann and Liberman, 1983; Remez *et al.*, 2001), but that accurate phonemic categorization was not necessary in order for this cross-language difference to occur. That being said, our argument about phonemic categorization rested on the results for a single stimulus series, and it is still possible that non-speech series exist that have cross-language discrimination differences. For example, Hay (2005) found some differences between Spanish and English speakers in their discrimination of a tone-onset non-speech analog of voice onset time (Pisoni, 1977), although with differences that were less clear than obtained with speech.

Instead of manipulating the speech/non-speech distinction, it is also possible to manipulate acoustic form to make stimuli that are acoustically unlike speech but remain intelligible. For example, Remez (1989) has demonstrated that sine-wave speech can be intelligible, and has used this finding to argue against the importance of auditory processing details in speech perception. In sine-wave speech, frequency modulated sinusoids are created that mimic natural formant movement and fricative center frequencies. Such a signal lacks the normal surface structure of speech (e.g., voicing and pitch variation) and has a highly unnatural quality, but it still can be understood. This implies that speech processing is based on some kind of higher-level representation of the signal, rather than on surface-level acoustic features. Remez has argued that the spectral-temporal variation of sine-wave and natural speech are perceived similarly because they are both related to speech articulations (cf. Hillenbrand *et al.*, 2011), and there is no purely auditory reason why either sine-wave or normal speech should be processed as a coherent whole (Remez *et al.*, 1994; cf. Barker and Cooke, 1999).

Other manipulations, such as noise or tone vocoding, can disrupt the natural acoustic form of speech while the speech remains intelligible (e.g., Shannon *et al.*, 1995). In this manipulation, speech is passed through a bank of bandpass filters and the amplitude within each band is used to modulate the amplitude of a corresponding carrier, such that the speech can take on the quality of the carrier (e.g., noise or inharmonic sinusoids) but preserve enough spectral dynamics for the speech to be understood. Although the recognition of vocoded speech is sometimes described as being based on temporal features (Shannon *et al.*, 1995), and sine-wave speech is described as being based on spectral modulations (e.g., Nittrouer *et al.*, 2009), they actually have very similar spectral-temporal variation, at least with enough channels in the vocoder. For example, sine-wave speech can be readily converted into intelligible noise-vocoded speech (e.g., Rosen *et al.*, 2011), indicating that the temporal information in sine-wave speech is not lost, and likewise the intelligibility of vocoded speech increases as spectral information is added (e.g., Shannon *et al.*, 1995).

The present study used manipulations both of acoustic form and the speech/non-speech distinction to explore how close stimuli need to acoustically resemble or sound subjectively like speech in order for cross-language discrimination differences to occur. The stimuli were all related to the English /r/-/l/ distinction, and were discriminated and identified by native speakers of English and Japanese. Experiment 1 used vocoders and sine-wave speech to create stimuli that sounded subjectively like speech but with unnatural acoustic forms, in order to test whether the English–Japanese discrimination difference for /r/ and /l/ depended on acoustic naturalness. If cross-language discrimination differences are reduced for stimuli that are acoustically unnatural but identifiable as /r/ and /l/, then this would suggest that the differences depend on the processing of surface-level acoustics rather than on phonological categorization. Experiment 2 instead used a series of vocoder manipulations that were designed to disrupt the categorization of stimuli while preserving the critical acoustic distinction between /r/ and /l/ (F3 formant differences), to examine whether it is possible to find stimuli that are difficult to categorize as /r/ and /l/, or even as speech, that are still discriminated differently by Japanese and English listeners. Such a result would likewise undermine the hypothesis that these differences are caused by phonological categorization.

## II. EXPERIMENT 1

The stimuli in this experiment were based on a synthesized /r/-/l/ continuum that modeled natural speech (Hattori and Iverson, 2009). They were passed through vocoders with different carriers (a harmonic carrier with a static F0, an inharmonic carrier with sinusoids that matched the filter bands, and a noise carrier), and the continuum was also converted to sine-wave speech. Natural recordings of /r/ and /l/ were processed the same way as a control. English and Japanese speakers performed a /r/-/l/ categorization task on the synthetic and natural speech, and discriminated stimuli at the /r/-/l/ boundary. Our aim was to assess whether the higher discrimination sensitivity that English speakers have at the /r/-/l/ boundary, compared to Japanese speakers, extend to intelligible speech with unnatural acoustic forms.

### A. Method

#### 1. Subjects

Twelve native southern British English speakers and 15 native Japanese speakers were tested. The ages ranged from 18 to 24 years (median = 21 years) for the British speakers, and 20 to 46 years (median = 29 years) for Japanese listeners. The Japanese listeners began learning English between 5 and 13 years of age (median = 13 years). None of the participants self-reported having hearing problems.

#### 2. Stimulus and apparatus

A synthesized /ra/-/la/ continuum interpolated between the best exemplars of English /r/ and /l/ found in a previous study (Hattori and Iverson, 2009), which had been based on a copy synthesis (Klatt and Klatt, 1990) of recordings from a female talker. There were a total of 76 stimulus steps, and the range of the stimuli were selected such that the English /r/-/l/ identification boundary found previously (Hattori and Iverson, 2009) occurred in the middle of the series. Respectively for /r/ and /l/, F3 varied from 2403 to 3508 Hz, the duration of the initial closure (i.e., before the transition to the vowel) varied from 31 to 96 ms, and the duration of the formant transition varied from 81 to 16 ms.

Natural recordings of eight /r/-/l/ initial-position minimal pair words were used to assess identification abilities (e.g., *row - low*). They were spoken by two speakers of southern British English (one female, one male). Natural recordings of ten BKB sentences (Bench *et al.*, 1979) were used to familiarize listeners with the acoustic transforms. They were all spoken by one female talker.

Vocoded versions of the natural and synthesized stimuli were created with either a harmonic carrier (*F0 Voc*, created with a 220 Hz pulse train), a noise carrier (*Noise Voc*), or with a bank of inharmonically related sinusoids matching the channel center frequencies (*Sine Voc*). See example spectrograms in Fig. 1. The center frequencies of the channels were selected so that they would be equally spaced with regard to the basilar membrane (Greenwood, 1990), but with a denser spacing of channels in the F3 region (2581 to 3431 Hz) to allow individuals to have discrimination thresholds comparable to the unprocessed speech. The center frequencies of the 20 channels were 137, 229, 348, 502, 704, 966, 1307, 1751, 2086, 2221, 2365, 2518, 2680, 2851, 3033, 3226, 3431, 4059, 5334, and 6992 Hz. For each stimulus, a low-pass filtered (300 Hz) amplitude envelope was calculated for each band, these amplitude envelopes were used to modulate the corresponding carrier in the vocoder, and carriers were re-filtered to attenuate side bands.

The sine-wave speech condition was created using Praat (Boersma and Weenink, 2010). The frequency values and amplitude envelopes for the first three formant transitions were measured from the synthetic stimuli automatically, then three sinusoids corresponding to the F1, F2, and F3 frequency and amplitude values were synthesized for each point on the continuum.

All stimuli were played over headphones (Sennheiser HD280) at a comfortable level in a sound-attenuated booth.

### B. Procedure

Listeners began the experiment with a short sentence identification task that was designed to familiarize them with each acoustic transform and demonstrate that they were able to hear each transform as speech. They heard a transformed version of a sentence, and were given ten sentence response options (e.g., *The old gloves were dirty*, *The house had nine rooms*). The interface was interactive; the natural version of a sentence was played when that response was clicked, such that subjects were able to click repeatedly on the response options, as well as replay the original transformed stimulus, until they felt ready to proceed to the next trial. Listeners heard each of the 10 sentences in each of the 4 acoustic transforms (i.e., 40 trials) presented in a random order.

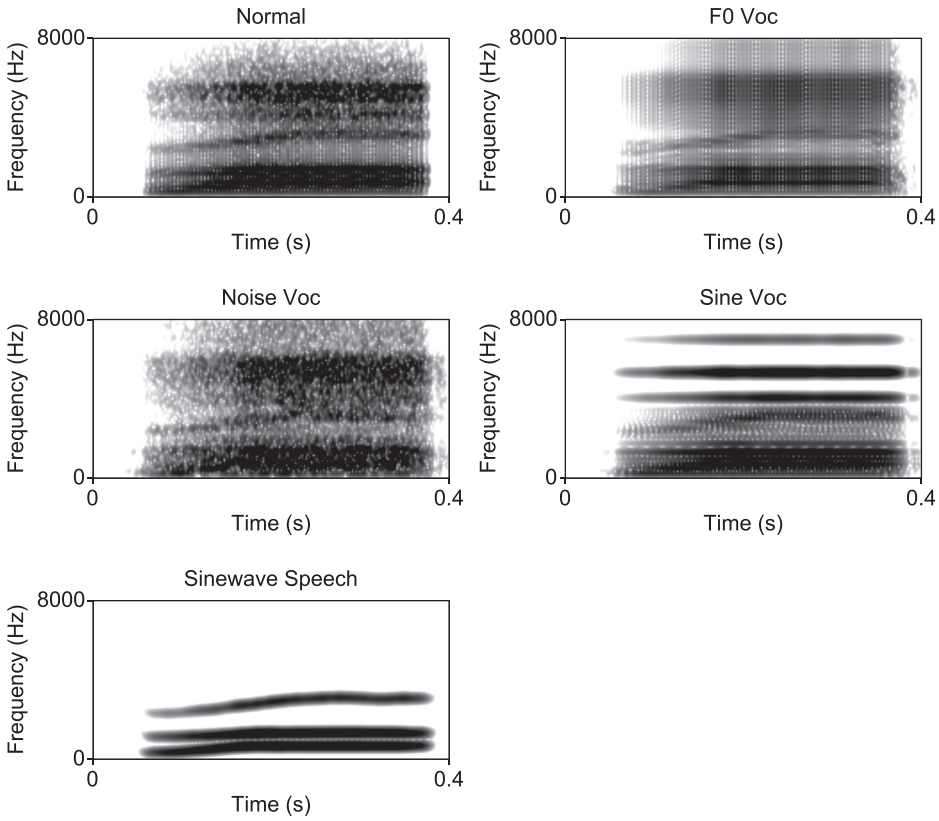


FIG. 1. Spectrograms of the /r/ stimulus endpoint under each of the acoustic transformations in Experiment 1. Normal is synthesized speech. F0 Voc, Noise Voc, and Sine Voc are vocoders with harmonic, noise, and inharmonic (sinusoids spaced according to basilar membrane distance) carriers.

Following this training, listeners were tested on forced-choice /r/-/l/ identification (i.e., hear one stimulus and decide whether it started with /r/ or /l/). The stimuli were recordings of minimal-pair words, presented normally (i.e., no acoustic processing) and under each acoustic transform. There were 80 trials presented in a random order (8 words $\times$ 2 talkers $\times$ 5 conditions).

Subjects performed a discrimination task that assessed their ability to discern acoustic differences that straddled the /r/-/l/ boundary. Listeners heard three stimuli on each trial (two same and one different) and had to choose the one that they thought was different. A modified Levitt procedure (Baker and Rosen, 2001) was used to adapt the acoustic difference between the stimuli to converge on the acoustic difference that yielded 71% correct performance on the task. The adaptive procedure used a one-up/two-down rule (i.e., one incorrect response made the acoustic difference larger, and two correct responses in a row made the acoustic difference smaller). Each adaptive series ended when there had been seven reversals (i.e., a change in whether the acoustic difference was becoming greater or smaller), or when the number of trials reached a maximum of 50. Discrimination thresholds were quantified as the mean acoustic difference on trials after the first three reversals. Listeners completed two adaptive series for each of the five stimulus conditions in a random order.

There are debates in the literature over which discrimination tasks best measure auditory processing for speech (e.g., Macmillan *et al.*, 1988; Pisoni and Lazarus, 1974). It has been argued that 4IAX provides a better measure of auditory sensitivity due to reduced memory demands (e.g., Pisoni and Lazarus, 1974), but the differences have mainly been found in roving designs (stimulus selection randomly drawn from a continuum), with the largest task-related differences found within categories where sensitivity is low. The adaptive procedure and oddity task used here limits these problems because the stimulus selection is effectively fixed after an initial period of convergence (i.e., focused around one point in the continuum, which reduces trace variance, e.g., Macmillan *et al.*, 1988), and the adaptive interval size keeps the acoustic differences above a behavioral threshold.

### C. Results and discussion

All subjects were extremely accurate in the training task; English subjects were all 100% correct and Japanese subjects averaged 98% correct at identifying the correct sentence from a selection of 10 different sentences. Although the task was designed to be relatively easy, their performance demonstrates that both language groups were able to interpret the novel acoustic transformations as speech.

Figure 2 displays the /r/-/l/ identification accuracy for each acoustic transformation and language group. In every condition, English speakers were mostly 100% correct, and individual scores from Japanese speakers ranged from chance to 100% correct. The various acoustic transformations appeared to have some effect on identification accuracy (e.g., the normal and sine vocoder conditions tended to have the highest performance), but performance was broadly similar. To test these differences, a logistic mixed model analysis was conducted with language and condition as fixed factors, and random factors of subject and stimulus with crossed intercepts. The analysis used the GLMM function in the R

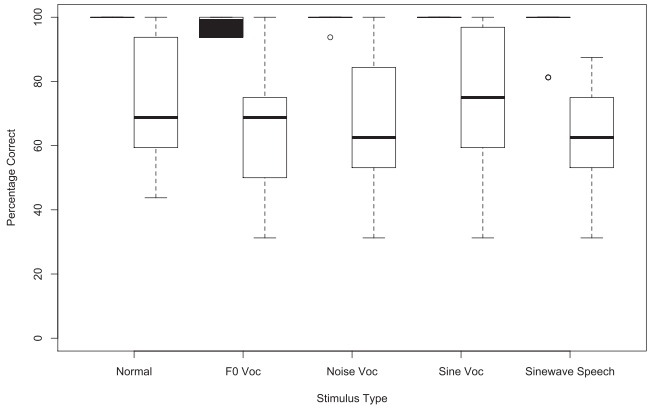


FIG. 2. Boxplots of percentage correct /r/-/l/ identification for Japanese (white) and English speakers (black, or a line at 100%). Boxplots display the range of data in terms of quartiles (e.g., second and third quartiles of the data distribution indicated by the boxes, with the lines in the boxes indicating medians). English speakers were significantly more accurate than Japanese speakers across stimuli with varying acoustic naturalness.

package lme4 (Bates *et al.*, 2015) and with a type II analysis-of-variance table calculated using the package CAR (Fox and Weisberg, 2011). Dummy coding was used for the independent variables. There was a significant main effect of language group, $\chi^2(1, N = 2160) = 50.33$, $p < 0.001$, and condition, $\chi^2(1, N = 2160) = 12.79$, $p = 0.012$, but no significant interaction, $p > 0.05$. The estimates and standard errors of the model were lang.J = −18.69 (20.64), cond.F0Voc = −15.54 (20.64), cond.NoiseVoc = −14.12 (20.64), cond.SineVoc = 0.05 (40.37), cond.SinewaveSpeech = −15.96 (20.65), Lang.J:cond.F0Voc = −15.20 (20.65), Lang.J:cond.NoiseVoc = −13.81 (20.65), Lang.J:cond.SineVoc = 0.04 (40.37), and Lang.J:cond.SinewaveSpeech = −15.41 (20.65).

Figure 3 displays the /r/-/l/ discrimination thresholds for each acoustic transformation and language group, which are expressed as a percentage of the entire continuum (e.g., 100% would mean that listeners discriminated continuum endpoints at 71% correct, and 10% meant that they needed the stimuli to be separated by 10% of the continuum range in order to discriminate at 71% correct). A linear mixed model

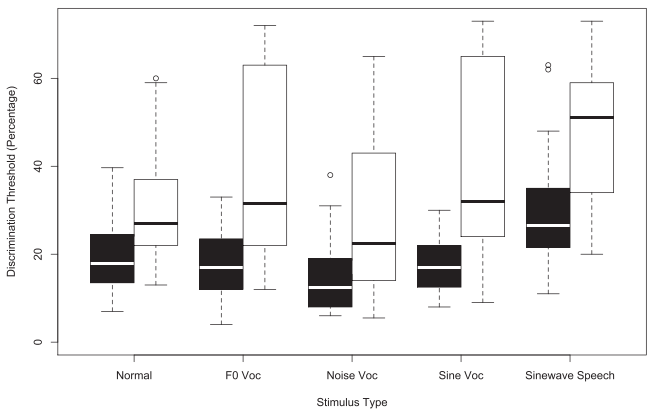


FIG. 3. Boxplots of 71% discrimination thresholds for Japanese (white) and English speakers (black). English speakers had significantly lower thresholds (i.e., better discrimination) than Japanese speakers across stimuli with varying acoustic naturalness.

analysis was conducted on the thresholds with language group and condition as fixed factors and with random intercepts for subject. The analysis was conducted using the lme function of the R package NLME (Pinheiro *et al.*, 2015), with the model being evaluated in an analysis-of-variance table. Dummy coding was used for language and condition. There were significant main effects of language group, $F(1,25) = 24.16$, $p < 0.001$, and condition, $F(4,100) = 16.08$, $p < 0.001$, but no significant interaction, $p > 0.05$. English speakers are thus more accurate than Japanese speakers at discriminating /r/ and /l/ stimuli regardless of the acoustic transform. The estimates and standard errors of the model were lang.J = 0.50 (0.17), cond.F0Voc = −0.10 (0.13), cond.NoiseVoc = −0.32 (0.13), cond.SineVoc = −0.09 (0.13), cond.SinewaveSpeech = 0.39 (0.13), Lang.J:cond.F0Voc = 0.29 (0.18), Lang.J:cond.NoiseVoc = 0.04 (0.18), Lang.J:cond.SineVoc = 0.25 (0.18), and Lang.J:cond.SinewaveSpeech = 0.02 (0.18).

Japanese speakers in this experiment had a wide range of identification performance, and we tested whether this was related to discrimination in a linear mixed model analysis on the Japanese speakers only, with their identification scores on normal stimuli added as a predictor. There was indeed a significant main effect of identification, $F(1,13) = 7.86$, $p = 0.015$, along with the main effect of condition found before, $F(4,52) = 8.92$, $p < 0.001$, but no significant interaction, $p > 0.05$. Japanese speakers who were better at identifying /r/ and /l/ thus had lower discrimination thresholds. The estimates and standard errors of the model were ID = −1.01 (0.52), cond.F0Voc = −0.08 (0.44), cond.NoiseVoc = −0.16 (0.44), cond.SineVoc = 0.53 (0.44), cond.SinewaveSpeech = 0.30 (0.44), ID:cond.F0Voc = 0.38 (0.59), ID:cond.NoiseVoc = −0.16 (0.59), ID:cond.SineVoc = −0.53 (0.59), and ID:cond.SinewaveSpeech = 0.15 (0.59).

Overall, the results demonstrate that cross-language differences for /r/ and /l/ occur for stimuli that are identifiable as those phonemes, even when their surface-level acoustic forms are unnatural. This implies that phoneme identification, or a speech-specific processing mode (e.g., Mann and Liberman, 1983; Remez *et al.*, 2001), may have driven these discrimination results.

## III. EXPERIMENT 2

Experiment 2 examined the role of categorization from a different perspective, by testing whether cross-language differences in discrimination can be found for non-speech analogs that do not sound like /r/ and /l/ but retain some of the acoustic characteristics of these phonemes.

One of the benefits of finding cross-language differences for vocoders in Experiment 1 is that it opens up a range of acoustic manipulations. Vocoders make it possible to arbitrarily manipulate the amplitude of different channels or swap in different carriers, whereas natural speech is harder to manipulate like this without producing artifacts. In the present experiment, the baseline, most natural, condition was the inharmonic vocoder of Experiment 1. At the other extreme, we created a condition designed to be similar to that of Miyawaki *et al.* (1975), using a vocoder with channels only

in the F3 range. For the continua in between, we used a variety of manipulations on the channels below F3 in order to make the entire stimulus sound more or less like speech (i.e., averaging the amplitude across adjacent channels to reduce spectral information, inverting the channel orders to disrupt spectral information, and combining noise and sinusoid carriers), but kept the critical F3 variation the same for all stimuli. We assessed how the patterns of acoustic variation in the channels below F3 affected categorization and speech likeness, in order to more fully explore the extent to which these factors are related to cross-language discrimination differences.

## A. Method

### 1. Subjects

Sixteen native southern British English speakers and 32 native Japanese speakers were tested, none of whom participated in Experiment 1. The ages ranged from 19 to 29 years (median = 22 years) for the British speakers and ranged from 20 to 43 years (median = 31 years) for Japanese listeners. The Japanese listeners began learning English between 10 and 13 years of age (median = 13 years). None of the participants self-reported having hearing problems.

### 2. Stimuli and apparatus

The stimuli were based on the synthetic continuum that was used in Experiment 1, using acoustic processing based on the sine vocoder of Experiment 1. The eight acoustic processing conditions were varied by manipulating the 8 channels below the F3 region (i.e., center frequencies 137–1751 Hz) and having the channels in the F3 region remain the same in all processing conditions. See Fig. 4 for example spectrograms. *Normal Voc* had the same channels as the sine vocoder in Experiment 1. *Four-channel Average* reduced spectral information by giving channels 1–4 the same envelope (an average of those 4 channels) and channels 5–8 the same envelope (an average of those 4 channels), which effectively reduced the spectral information below F3 to two broader channels. *Eight-channel Average* similarly averaged the envelopes across channels 1–8, effectively reducing the spectral information to 1 broad channel. *Eight-channel Average Noise* did the same averaging, but used a noise carrier for channels 1–8 and kept the sinusoidal carriers for the

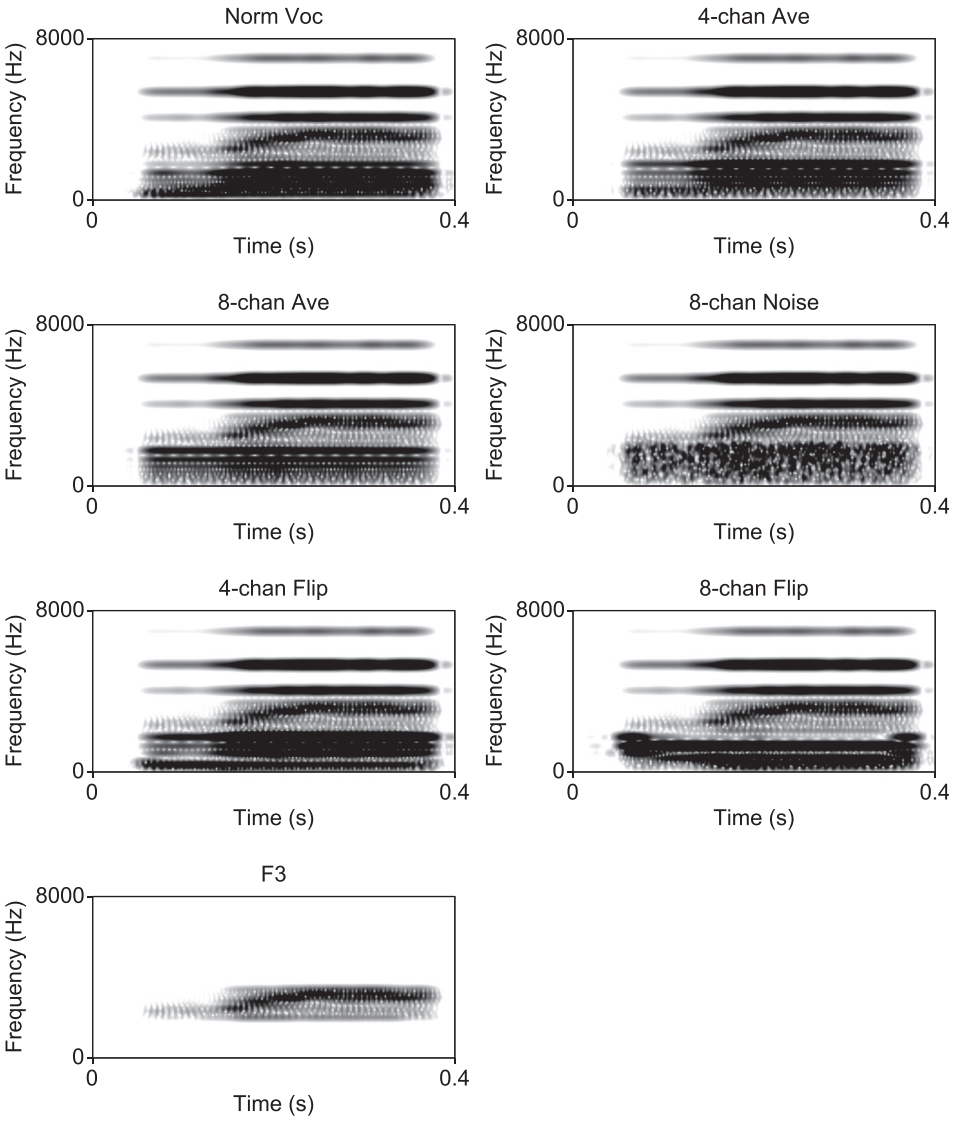


FIG. 4. Spectrograms of the /r/ stimulus endpoint under each of the acoustic transformations in Experiment 1. The F3 region was the same for all stimuli. Norm Voc was the sine vocoder from Experiment 1. Four- and eight-chan Ave reduced spectral resolution by averaging the amplitude envelopes of adjacent carriers below the F3 region. Eight-chan noise was the same as Eight-chan Ave, except that a noise carrier was used for the lower channels. Four- and Eight-chan Flip disrupted the spectra by inverting the channel orders below F3. F3 had energy in the F3 frequency region only.

higher channels. *Four-channel Flip* inverted the spectral channels in sets of four. That is, the amplitude information in channels 1–4 were delivered to channels 4–1, respectively (e.g., the amplitude envelope derived from the lowest band was used to modulate a sinusoid that matched the fourth highest band), the analogous transformation was conducted for channels 5–8, and the higher-channels remained unchanged. *Eight-channel Flip* similarly inverted channels 1–8. *F3* only presented information in the F3 range.

The apparatus was the same as in Experiment 1.

### B. Procedure

Listeners began the experiment with a short sentence identification task that was the same as the familiarization task in Experiment 1, except that listeners were exposed to only two acoustic transforms, the Sine Voc and the Noise Voc transformations used in Experiment 1. This task thus familiarized listeners with speech created with inharmonic sinusoid and noise carriers, but they were not given the transformations used in the present study that were designed to make speech less intelligible.

Following this training, listeners were given a forced-choice consonant identification task with 15 alternatives (*sa*, *za*, *tha*, *fa*, *va*, *sha*, *ja*, *cha*, *ya*, *ra*, *la*, *ma*, *na*, *wa*, and *noise*). This task was designed to see whether listeners could interpret the stimuli as speech at all (they gave the noise response if they could not), and also see whether the stimuli could be distinguished from other consonants (i.e., beyond the basic /r/-/l/ distinction). Listeners heard the /r/-/l/ continua endpoints under each of the 7 stimulus conditions (i.e., 14 stimuli, presented 4 times each in a random order), and these were combined with an equal number of natural consonant stimuli processed with the Sine Voc transformation (i.e., one natural stimulus for each of the response alternatives, excluding the noise response, presented four times each). The additional stimuli that matched the other responses were included so that subjects would be less biased to respond with /r/ and /l/, and subjects had not been told that this study was focused on these consonants.

Listeners performed a discrimination task that was identical to the adaptive procedure used in Experiment 1 (i.e., two adaptive series for each of the seven stimulus conditions, presented in a random order).

At the end of the session, listeners were tested on forced-choice /r/-/l/ identification (i.e., hear one stimulus and decide whether it started with /r/ or /l/). This was also the same as in Experiment 1, except that listeners were tested only on normal speech, and the Sine Voc and Noise Voc transformations from Experiment 1. This identification test was designed to verify that listeners could hear /r/ and /l/ differences with the vocoders that had been used in the initial familiarization, but it was presented at the end of the experiment so that it would be less transparent to the subjects that the experiment was focused on /r/ and /l/.

### C. Results and discussion

Similar to the results of Experiment 1, all subjects had little difficulty with the sentence training task; all subjects were perfect except for one Japanese listener who made one error. The /r/-/l/ identification for natural stimuli under these conditions (i.e., final identification task with sine and noise vocoders) was very accurate for English subjects and varied widely for Japanese subjects. To test these differences, a logistic mixed model analysis was conducted with language and condition as fixed factors, and random factors of subject and stimulus with crossed intercepts. The analysis used the GLMM function in the R package lme4 (Bates *et al.*, 2015) and with a type II analysis-of-variance table calculated using the package CAR (Fox and Weisberg, 2011). Dummy coding was used for language and condition. There was a significant main effect of language group, $\chi^2(1,\ N=4608)=38.60$, $p<0.001$, but no significant main effect of condition or an interaction, $p>0.05$. Thus, both groups of listeners were able to perceive these types of transformations as speech, but Japanese speakers had more difficulty distinguishing /r/ and /l/. The estimates and standard errors of the model were lang.J = −18.83 (13.35), cond.Normal = −0.28 (20.64), cond.SineVoc = 13.89 (13.35), Lang.J:cond.Normal = −0.01 (23.18), and Lang.J:cond.SineVoc = −13.89 (13.35).

Figure 5 displays the identification results for stimulus endpoints under the experimental conditions, coded in terms of the proportion of correct responses (i.e., labeling the appropriate endpoint as /r/ or /l/), incorrect phoneme responses, and the noise response (i.e., non-speech). The transformations were successful in creating a set of stimuli that varied both in terms of how accurately they could be identified and how often they sounded like speech at all. For English speakers, all continua except Normal Voc and Four-channel Average had endpoints that were identified correctly on less than half of the trials, and two continua (Eight-

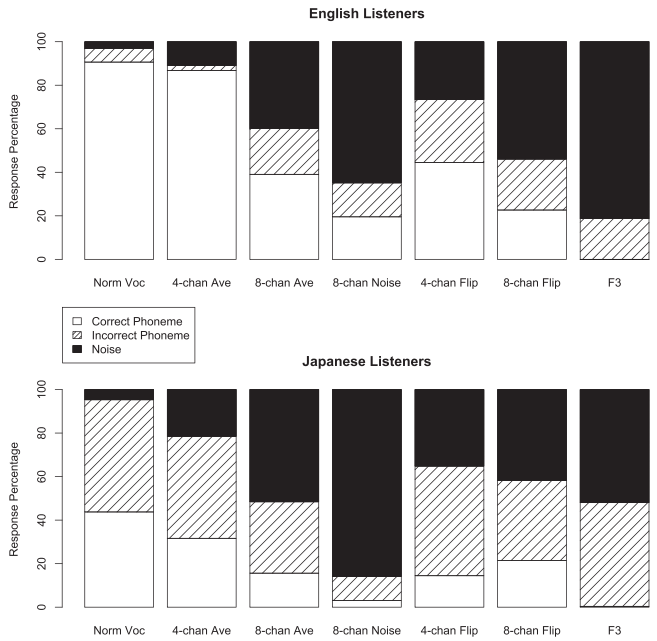


FIG. 5. Stacked bar charts of identification percentages for English and Japanese speakers, scored in terms of whether the /r/ and /l/ phoneme was identified correctly, as an incorrect phoneme, or was judged to not sound like speech (i.e., a noise). English speakers were more accurate than were Japanese, but both groups found the stimuli harder to identify and less like speech as the stimuli became acoustically further from normal speech.

channel Noise and F3) were perceived as non-speech on more than half of the trials. The results were similar for Japanese speakers, although there were more incorrect responses owing to their /r/-/l/ identification difficulty. Chi-square tests revealed that there were significant cross-language differences in the distribution of responses (coded into three categories) for each stimulus continuum [Norm Voc: $\chi^2(1, N = 384) = 80.14$, $p < 0.001$; 4-chan Ave: $\chi^2(1, N = 384) = 109.89$, $p < 0.001$; 8-chan Ave: $\chi^2(1, N = 384) = 26.51$, $p < 0.001$; 8-chan Noise: $\chi^2(1, N = 384) = 33.04$, $p < 0.001$; 4-chan Flip: $\chi^2(1, N = 384) = 42.60$, $p < 0.001$; 8-chan Flip: $\chi^2(1, N = 384) = 7.45$, $p = 0.024$; F3: $\chi^2(1, N = 384) = 31.12$, $p < 0.001$].

Figure 6 displays the /r/-/l/ discrimination thresholds for each acoustic transformation and language group, which are expressed as a percentage of the entire continuum. A linear mixed model analysis was conducted on the thresholds with language group and condition as fixed factors and with random intercepts for subject. The analysis was conducted using the lme function of the R package NLME (Pinheiro *et al.*, 2015), with the model being evaluated in an analysis-of-variance table. There were significant main effects of language group, $F(1,46) = 13.43$, $p < 0.001$, and condition, $F(6,275) = 7.44$, $p < 0.001$, as well as a significant interaction of these factors, $F(6,275) = 5.82$, $p < 0.001$. Reverse Helmert coding was used for condition, with the stimuli ordered the same way as in the figure based on the amount of spectral disruption made to the lower channels, so that the linear mixed model compared each condition to the sum of the data for the previous conditions. The contrasts demonstrated that the language effect was greatest for the first two continua that were accurately identified as /r/ and /l/, significantly reduced for the middle three continua (Eight-channel Average, Eight-channel Noise, Four-channel Flip), and significantly reduced again for the last two continua (Eight-channel Flip, F3). Likewise, independent-sample *t*-tests demonstrated that there was no significant effect of language for the last two continua, but English and Japanese speakers were significantly different for all other continua, $p < 0.05$. Cross-language discrimination differences were thus found for the middle range of stimuli that were not accurately identified as /r/ and /l/, and sometimes did not sound like speech. The estimates and standard errors of the model were lang.J = 0.27 (0.07), cond.4-chanAve = −0.16 (0.06), cond.8-chanAve = 0.15 (0.35), cond.8-chanNoise = −0.06 (0.02), cond.4-chanFlip = 0.02 (0.02), cond.8-chanFlip = −0.04 (0.02), cond.F3 = 0.06 (0.01), Lang.J:cond.4-chanAve = 0.23 (0.07), Lang.J:cond.8-chanAve = −0.10 (0.04), Lang.J:cond.8-chanNoise = −0.04 (0.03), Lang.J:cond.4-chanFlip = −0.02 (0.02), Lang.J:cond.8-chanFlip = −0.06 (0.02), and Lang.J:cond.F3 = −0.04 (0.02).

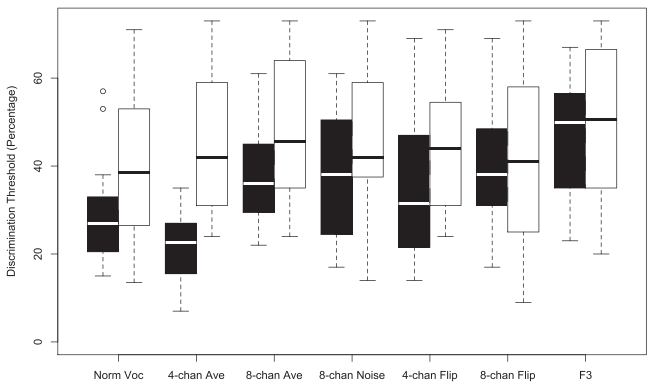


FIG. 6. Boxplots of discrimination thresholds for Japanese (white) and English (black) speakers. There were cross-language differences for the stimuli that sounded most like /r/ and /l/ (Norm Voc and 4-chan Ave), and no significant differences for the stimuli with the largest spectral disruption of the lower channels (8-chan Flip and F3). However, there were also cross-language differences for stimuli in the middle that were hard to identify and often sounded unlike speech, indicating an effect of language experience on general perceptual processing.

Figure 7 displays the mean discrimination results for each condition, for English and Japanese speakers, plotted against the mean noise and speech responses. Although the proportion of noise responses was significantly related to discrimination thresholds for English speakers, $r = 0.89$, $p = 0.008$, it did not reach that level for Japanese speakers, $r = 0.67$, $p = 0.082$. Also, both groups had similar proportions of noise responses, so the extent that the stimuli sounded like non-speech cannot account for the cross-language discrimination differences. For the correct responses, the correlation was significant for English, $r = -0.93$, $p = 0.002$, and Japanese speakers, $r = -0.81$, $p = 0.03$. Moreover, on this measure the language groups fell closer onto a single line. Thus, one cannot discard the hypothesis that phonological categorization contributed to this pattern of discrimination results, even though the cross-language differences extended to stimuli that were not often given a correct phonological label.

Japanese speakers in this experiment had a wide range of identification performance, and we tested whether this was related to discrimination in a linear mixed model analysis on the Japanese speakers only, with their identification scores on normal stimuli added as a predictor. There was a significant main effect of identification, $F(1,30) = 4.57$, $p = 0.041$, along with the main effect of condition found before, $F(6,179) = 2.72$, $p = 0.015$, but no significant interaction, $p > 0.05$. Japanese speakers who were better at identifying /r/ and /l/ thus had lower discrimination thresholds, further implicating the role of phonological encoding. The estimates and standard errors of the model were ID = −0.53 (0.25), cond.4-chanAve = 0.26 (0.20), cond.8-chanAve = −0.22 (0.12), cond.8-chanNoise = 0.00 (0.08), cond.4-chanFlip = −0.01 (0.06), cond.8-chanFlip = −0.10 (0.05), cond.F3 = 0.00 (0.04), ID:cond.4-chanAve = −0.24 (0.27), ID:cond.8-chanAve = 0.37 (0.15), ID:cond.8-chanNoise = 0.03 (0.11), ID:cond.4-chanFlip = 0.01 (0.08), ID:cond.8-chanFlip = 0.10 (0.07), and ID:cond.F3 = 0.02 (0.06).

## IV. GENERAL DISCUSSION

Our results suggest that language experience affects speech-specific processing for /r/ and /l/, in that cross-language differences were found for stimuli with unnatural surface-level acoustics (Experiment 1), the discrimination differences were strongly related to how accurately the phonemes were labeled (Experiment 2), and they disappeared

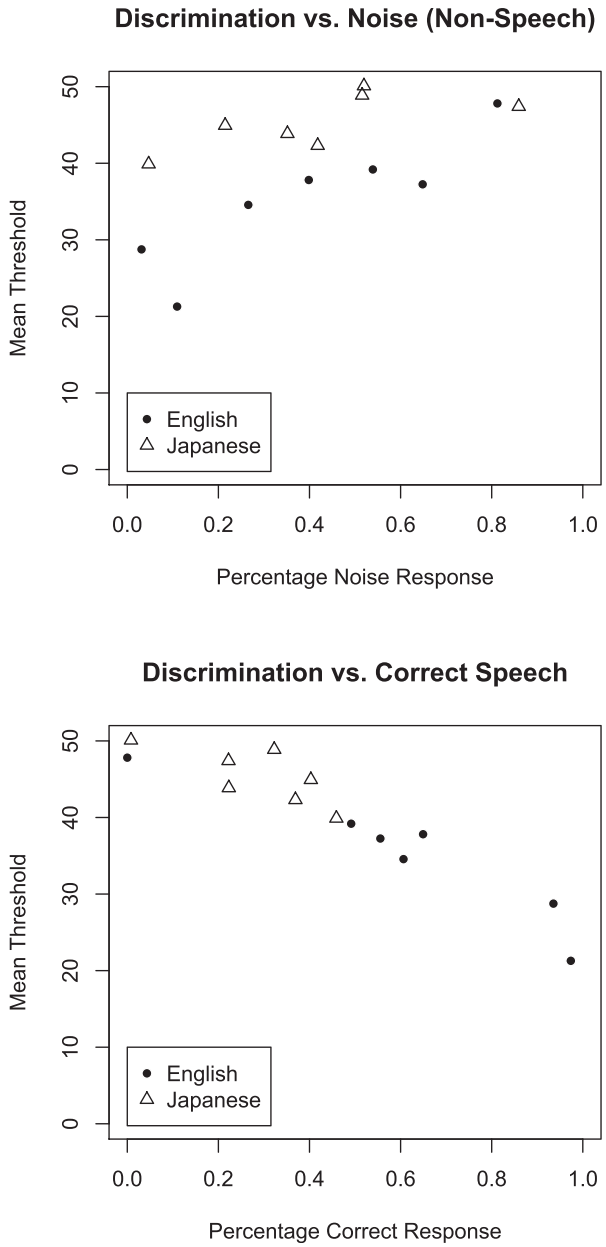


FIG. 7. Scatterplots of mean discrimination performance vs the mean noise and correct identification responses, for each language group and continuum. The extent that the stimuli sounded like non-speech (noise) did not account for the cross-language differences in discrimination, but there was a strong relationship between phonological identification and discrimination performance across the range of continua.

for stimuli that had spectral-temporal variation that was different enough from normal speech (Experiment 2). That being said, Experiment 2 adds to our previous examples (Iverson *et al.*, 2011) of stimuli that are difficult to categorize phonemically, and sometimes not even sound like speech, but are affected by language experience.

The results replicate the findings of Miyawaki *et al.* (1975) of no cross-language differences for F3 transitions only, but clarify that cross-language differences can be found for other sounds that do not sound consistently like speech. Our results also fit with Hay (2005), who found significant, but diminished, Spanish–English discrimination differences for a tone-onset analog of voice onset time. We think that their non-speech analog is comparable to our middle range of stimuli in Experiment 2 (i.e., close enough to speech to produce cross-language effects but not close enough to sound like speech), and predict that the cross-language difference would be eliminated for stimuli that were less like speech in terms of spectral-temporal properties (e.g., if the tones were moved outside of the range of F1 and F2), and increased for stimuli that were more like speech. In addition, the results are compatible with Krishnan *et al.* (2005, 2009a, 2009b), who found cross-language differences in the processing of Mandarin tone analogs that did not have a normal speech spectrum, but only when the pitch contours were similar to Mandarin tones.

Xu *et al.* (2006; cf. Bent *et al.*, 2006) similarly found cross-language discrimination differences for Mandarin and English speakers listening to Mandarin tones and a non-speech harmonic complex with a pitch contour identical to the tone stimuli. However, rather than attributing the non-speech differences to auditory processing, they hypothesized that they were caused by cross-language differences in short-term categorical memory; listeners may form domain-general memory representations based on an initial feature analysis of the sound, and it is possible that this feature analysis is affected by long-term experience with speech even when the input does not sound like speech. In the present study, we cannot distinguish between general auditory processing and general short-term memory processes. That is, it is possible that even when stimuli did not sound consistently as /r/ and /l/ categories, listeners may have covertly labeled the stimuli in a way that mirrored how they label speech sounds, and this labeling may have produced cross-language differences. This explanation has some plausibility, given that the stimulus continua in the middle (i.e., 8-chan Ave, 8-chan Noise, 4-chan Flip) had some acoustic similarity to speech and these stimuli received correct phoneme identifications from English speakers on a minority of trials. That being said, our results differ from Xu *et al.*, in that Mandarin speakers discriminated the speech and non-speech tones nearly the same, whereas in the present study discrimination thresholds clearly increase as the stimuli become further from speech and are less reliably labeled as the correct phonemes; listeners could not have been applying the same encoding strategy across all of the continua.

The present results contribute to the debate about whether phonetic processing reduces sensitivity within a category (*acquired similarity*), increases sensitivity at a boundary (*acquired distinctiveness*), or both (e.g., Iverson and Kuhl, 2000; Liberman *et al.*, 1961). The present study measured sensitivity at the boundary and the results strongly support acquired distinctiveness; discrimination thresholds were lower when the stimuli most resembled speech in Experiment 2, even though the acoustic differences were the same. The difference is that it is not clear to what extent the acquired distinctiveness at the category boundary in the present study was a direct result of phonological encoding or was caused by auditory/phonetic processing.

In the end, we are left with a view of language experience affecting perceptual processes that can be considered to be both general (i.e., not requiring a stimulus that sounds

like speech) and speech-specific (i.e., only affecting stimuli that have a spectral-temporal similarity to speech, and related strongly to phonological labeling). Stimuli that are in the gray area between speech and non-speech seem to produce discrimination results that likewise have multiple interpretations. It is thus questionable whether a sharp division of general auditory and speech-specific processing modes exists (e.g., Liberman and Mattingly, 1989). There are likely effects of language experience at many levels of perceptual and cognitive processing, and some of these effects may be speech-specific only in the sense that they do not extend very far to acoustically dissimilar sounds.

## ACKNOWLEDGMENTS

This research was supported by a Wellcome Trust grant awarded to P.I. We are grateful to Melanie Pinet and Yasuaki Shinohara for helping with data collection.

Andruski, J. E., Blumstein, S. E., and Burton, M. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**, 163–187.
Baker, R. J., and Rosen, S. (**2001**). "Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking," Br. J. Audiol. **35**, 43–52.
Barker, J., and Cooke, M. P. (**1999**). "Is the sine-wave speech cocktail party worth attending?," Speech Commun. **27**, 159–174.
Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-9," https://CRAN.R-project.org/package=lme4.
Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.
Bent, T., Bradlow, A. R., and Wright, B. A. (**2006**). "The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds," J. Exp. Psychol.: Human Percept. Perform. **32**, 97–103.
Best, C. T. (**1995**). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Baltimore, MD), pp. 171–204.
Boersma, P., and Weenink, D. (**2010**). "Praat: Doing phonetics by computer" [Computer program]. Retrieved from http://www.praat.org/ (Last viewed June 1, 2013).
Diehl, R. L., and Kluender, K. (**1989**). "On the objects of speech perception," Ecol. Psychol. **1**, 121–144.
Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Baltimore, MD), pp. 233–277.
Fox, J., and Weisberg, S. (**2011**). *An {R} Companion to Applied Regression*, 2nd ed. (Sage, Thousand Oaks, CA).
Goto, H. (**1971**). "Auditory perception by normal Japanese adults of the sounds 'L' and 'R,' " Neuropsychologia **9**, 317–323.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Hattori, K., and Iverson, P. (**2009**). "English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy," J. Acoust. Soc. Am. **125**, 469–479.
Hay, J. F. (**2005**). "How auditory discontinuities and linguistic experience affect the perception of speech and non-speech in English- and Spanish-speaking listeners," Ph.D. thesis, University of Texas at Austin.
Hillenbrand, J. M., Clark, M. J., and Baer, C. A. (**2011**). "Perception of sine-wave vowels," J. Acoust. Soc. Am. **129**, 3991–4000.
Iverson, P., Ekanayake, D., Hamann, S., Sennema, A., and Evans, B. G. (**2008**). "Category and perceptual interference in second-language phoneme learning: An examination of English /w/-/v/ learning by Sinhala, German, and Dutch speakers," J. Exp. Psychol.: Human Percept. Perform. **34**, 1305–1316.
Iverson, P., and Kuhl, P. K. (**2000**). "Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism?," Percept. Psychophys. **62**, 874–886.
Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (**2003**). "A perceptual interference account of acquisition difficulties for non-native phonemes," Cognition **87**, B47–B57.
Iverson, P., Wagner, A., Pinet, M., and Rosen, S. (**2011**). "Cross-language specialization in phonetic processing: English and Hindi perception of /w/-/v/ speech and non-speech," J. Acoust. Soc. Am. **130**, EL297–EL303.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis and perception of voice quality variations among male and female talkers," J. Acoust. Soc. Am. **87**, 820–856.
Krishnan, A., Gandour, J. T., Bidelman, G. M., and Swaminathan, J. (**2009a**). "Experience dependent neural representation of dynamic pitch in the brainstem," NeuroReport **20**, 408–413.
Krishnan, A., Swaminathan, J., and Gandour, J. T. (**2009b**). "Experience dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context," J. Cognit. Neurosci. **21**, 1092–1105.
Krishnan, A., Xu, Y., Gandour, J., and Cariani, P. (**2005**). "Encoding of pitch in the human brainstem is sensitive to language experience," Cognit. Brain Res. **25**, 161–168.
Kuhl, P. K., and Iverson, P. (**1995**). "Linguistic experience and the 'perceptual magnet effect,' " in *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (York Press, Baltimore, MD), pp. 121–154.
Liberman, A. M., Harris, K. S., Kinney, J. A., and H. Lane. (**1961**). "The discrimination of relative onset-time of the components of certain speech and non-speech patterns," J. Exp. Psychol. **61**, 379–388.
Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.
Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (**1988**). "Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am. **84**, 1262–1280.
Mann, V. A., and Liberman, A. M. (**1983**). "Some differences between phonetic and auditory modes of perception," Cognition **14**, 211–235.
McMurray, B., Tanenhaus, M. K., and Aslin, R. A. (**2002**). "Gradient effects of within-category phonetic variation on lexical access," Cognition **86**, B33–B42.
Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (**1975**). "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," Percept. Psychophys. **18**, 331–340.
Nittrouer, S., Lowenstein, J. H., and Packer, R. R. (**2009**). "Children discover the spectral skeletons in their native language before the amplitude envelopes," J. Exp. Psychol.: Human Percept. Perform. **35**, 1245–1253.
Norris, D., and McQueen, J. M. (**2008**). "Shortlist B: A Bayesian model of continuous speech recognition," Psychol. Rev. **115**, 357–395.
Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**, 299–325.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (**2015**). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-122, http://CRAN.R-project.org/package=nlme.
Pisoni, D. B. (**1977**). "Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stop consonants," J. Acoust. Soc. Am. **61**, 1352–1361.
Pisoni, D. B., and Lazarus, J. H. (**1974**). "Categorical and noncategorical modes of speech perception along the voicing continuum," J. Acoust. Soc. Am. **55**, 328–333.
Remez, R. E. (**1989**). "When the objects of perception are spoken," Ecol. Psychol. **1**, 161–180.
Remez, R. E., Pardo, J. S., Piorkowski, R. L., and Rubin, P. E. (**2001**). "On the bistability of sinewave analogs of speech," Psychol. Sci. **12**, 24–29.
Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (**1994**). "On the perceptual organization of speech," Psychol. Rev. **101**, 129–156.
Repp, B. H. (**1984**). "Categorical perception: Issues, methods, findings," in *Speech and Language: Advances in Basic Research and Practice*, Vol. 10, edited by N. J. Lass (Academic Press, Orlando, FL), pp. 244–322.

Rosen, S., and Iverson, P. (**2007**). "Constructing adequate non-speech analogues: What is special about speech anyway?," Develop. Sci. **10**, 165–168.
Rosen, S., Wise, R. J. S., Chadha, S., Conway, E.-J., and Scott, S. K. (**2011**). "Hemispheric asymmetries in speech perception: Sense, nonsense and modulations," PLoS One **6**, e24672.
Schouten, B., Gerrits, E., and Van Hessen, A. (**2003**). "The end of categorical perception as we know it," Speech Commun. **41**, 71–80.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.
Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (**1970**). "Motor theory of speech perception: A reply to Lane's critical review," Psychol. Rev. **77**, 234–249.
Toscano, J. C., McMurray, B., Dennhardt, J., and Luck, S. J. (**2010**). "Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech," Psychol. Sci. **21**, 1532–1540.
Trubetzkoy, N. S. (**1969**). *Principles of Phonology* (C. A. M. Baltaxe, Trans.) (University of California Press, Berkeley) (Original work published in 1939).
Werker, J. F., and Curtin, S. (**2005**). "PRIMIR: A developmental framework of infant speech processing," Lang. Learn. Devel. **1**, 197–234.
Whalen, D. H. (**1997**). "What duplex perception tells us about speech perception," in *Papers from the Panels, CLS 33*, edited by K. Singer, R. Eggert, and G. Anderson (Chicago Linguistic Society, Chicago, IL), pp. 435–446.
Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**, 1063–1074.