Bilingualism: Language and Cognition: page 1 of 20 © Cambridge University Press 2011 doi:10.1017/S1366728911000447

Can native Japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency?[*]

ERIN M. INGVALSON
Carnegie Mellon University
LORI L. HOLT
Carnegie Mellon University
JAMES L. McCLELLAND
Stanford University

(Received: December 9, 2009; final revision received: July 9, 2011; accepted: July 28, 2011)

Many attempts have been made to teach native Japanese listeners to perceptually differentiate English /r–l/ (e.g. rock–lock*). Though improvement is evident, in no case is final performance native English-like. We focused our training on the third formant onset frequency, shown to be the most reliable indicator of /r–l/ category membership. We first presented listeners with instances of synthetic /r–l/ stimuli varying only in F3 onset frequency, in a forced-choice identification training task with feedback. Evidence of learning was limited. The second experiment utilized an adaptive paradigm beginning with non-speech stimuli consisting only of /r/ and /l/ F3 frequency trajectories progressing to synthetic speech instances of /ra–la/; half of the trainees received feedback. Improvement was shown by some listeners, suggesting some enhancement of /r–l/ identification is possible following training with only F3 onset frequency. However, only a subset of these listeners showed signs of generalization of the training effect beyond the trained synthetic context.*

Keywords: /r–l/, second language speech perception, training

Learning a new language in adulthood can present challenges. One challenge that often arises is learning to perceive and produce the new language's sounds. A well-studied example is the difficulty native Japanese (NJ) speakers have with the English sounds /r/ as in *rock* and /l/ as in *lock*. Theories about the source of this difficulty vary (e.g., Flege, 2002; Kuhl, 1993; Lenneberg, 1967), furthering interest in this topic as a means of identifying constraints on adult language learning (Flege, Takagi & Mann, 1996; Guion, Flege, Akahane-Yamada & Pruitt, 2000; Takagi & Mann, 1995).

Additionally, a considerable amount of effort has been directed at targeted interventions that aim to teach participants to differentiate non-native contrasts reliably (e.g., Jamieson & Morosan, 1986; Strange & Dittman, 1984). These studies, while demonstrating that improvement is possible, have also served to highlight the difficulty NJ listeners have with English /r–l/ (Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997; Iverson, Hazan & Bannister, 2005; Lively, Logan & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura & Yamada, 1994; Logan, Lively & Pisoni, 1991; McCandliss, Fiez, Protopapas, Conway & McClelland, 2002).

Strange and Dittman (1984) made an early attempt to train NJ listeners to distinguish /r–l/. Their stimuli were synthetic items from *rake–lake* and *rock–lock* continua (MacKain, Best & Strange, 1982). NJ listeners were trained to discriminate stimuli drawn from one of the two continua, then were tested on their ability to discriminate pairs from both continua. Listeners showed evidence of learning via improved discrimination on both continua but failed to reliably discriminate untrained natural speech /r–l/ minimal pairs (e.g., *right–light*).

McCandliss et al. (2002) also used synthetic stimuli, but these were instances of *rock–lock* and *road–load* produced by one male native English (NE) speaker that were modified to emphasize the initial contrast. Participants in this task were trained to identify stimuli on one continuum then tested on their ability to identify and discriminate stimuli from both continua. The NJ listeners in this study better identified and discriminated both the trained and untrained continua at the post-test. Generalization to natural speech was not assessed, but improvement is unlikely (see Strange & Dittman, 1984).

The improvement seen in the above training studies has generally not been viewed as truly general speech perception learning given that the NJ listeners (in cases

[*] We wish to thank Daniel Dickison for serving as translator and interpreter and Robert Kass for numerous statistical consultations, particularly the suggestion of Fisher's combined probability test. We also wish to thank several anonymous reviewers for their helpful comments. Portions of this work were presented at the 2003 meeting of the Psychonomic Society and the 2005 meeting of the Cognitive Science Society. This work was supported by NIH grant 3R01-DC004674-06S1, NSF grant BCS-0746067, and a grant from The Bank of Sweden Tercentenary Foundation to the second author and NIMH grant P50-MH64445 to the third author.

Address for correspondence:
Erin Ingvalson, Department of Communication Sciences and Disorders, Northwestern University, 2240 Campus Dr., Evanston, IL 60208, USA
ingvalson@northwestern.edu

where this was tested) were unable to reliably differentiate natural speech /r–l/ minimal pairs. One response to this has been to raise the possibility that NJ listeners would be better able to learn the characteristics of English /r–l/ categories via training using natural speech /r–l/. If these natural speech stimuli were produced by a variety of NE speakers in a large number of contexts, the greater acoustic variability might enable the NJ listeners to learn those acoustic properties that reliably differentiate /r–l/. This approach was pursued by Pisoni and colleagues (Bradlow et al., 1999; Bradlow et al., 1997; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991). The stimuli in their studies were naturally produced instances of /r–l/ minimal pairs spoken by several NE speakers, both male and female. NJ listeners were trained to identify the words with feedback. Testing occurred via identification of trained and untrained tokens produced by talkers used in training and talkers not used in training (both types of talkers produced instances of trained and untrained tokens). Unlike the Strange and Dittman (1984) study, NJ listeners in these studies showed both improvement on trained materials and generalization to untrained natural speech stimuli – both untrained words produced by talkers used in training and untrained words produced by talkers not used in training. However, even after training their identification performance still fell well below NE levels.

One possibility for these incomplete success stories is an inherent limitation in the ability of adult NJ listeners to learn the distinction (Takagi, 2002; Takagi & Mann, 1995), possibly reflecting a broad, age-dependent cessation of plasticity for this aspect of language learning (Johnson & Newport, 1989). While certainly this possibility is consistent with results to date, there remains an alternative: the reason for the incomplete success may lie in the fact that the cues NJ speakers learn to utilize may not be the crucial cue NE speakers use to differentiate /r–l/. In studies using synthetic speech (McCandliss et al., 2002; Strange & Dittman, 1984) listeners may learn to rely on cues that distinguish the training stimuli but are not robust cues to the /r–l/ contrast across the full range of natural /r–l/. This would account for high levels of performance on the training stimuli but poor generalization to natural speech. In studies using natural speech from a range of speakers (Lively et al., 1991; Logan et al., 1993; Logan et al., 1994) there may be a similar difficulty. NJ participants may learn to rely on a variety of partial cues that weakly covary with the /r–l/ contrast but which are nevertheless imperfect cues to the /r–l/ distinction. This would explain why trained participants show a real and persistent generalizable learning effect (Bradlow et al., 1997; Bradlow et al., 1999), but where final attainment is non-native-like. Similar to natural-speech training, long-term immersion may also result in reliance on a variety of partial cues, explaining why NJ speakers with extensive immersion experience with English also show improved but non-native levels of performance (Aoyama, Flege, Guion, Akahane-Yamada & Yamada, 2004; Gordon, Keyes & Young, 2001).

There are several acoustic cues to /r–l/ category membership, the most well documented being the onset frequency of the third formant, F3, and the closure and transition duration of the first formant, F1, with F3 onset frequency being the most consistently reliable indicator of category membership (Espy-Wilson, 1992; O’Connor, Gertsman, Liberman, Delattre & Cooper, 1957). Instances of /r/ are typified by F3 onsets below the vowel steady state and long F1 closures followed by short transitions to the vowel steady state; instances of /l/ are typified by F3 onsets equal to or above the vowel steady state and short F1 closures followed by longer transitions to the vowel steady state. NE speakers make use of both of these cues when perceiving /r–l/ (Gordon et al., 2001; O’Connor et al., 1957; Polka & Strange, 1985; Underbakke, Polka, Gottfried & Strange, 1988). However, NE listeners place the greatest weight on F3 and changes in F3 alone are sufficient to shift NE listeners’ responses from /r/ to /l/ (Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Ketterman & Siebert, 2003; Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975; O’Connor et al., 1957; Yamada & Tohkura, 1990). NE speakers also weight F3 heavily in production, emphasizing this cue’s importance (Lotto, Sato & Diehl, 2004).

Conversely, NJ listeners appear to rely more heavily on less reliable cues, most notably the onset frequency of the second formant, F2, in both perception and production (Iverson et al., 2003; Lotto et al., 2004; Yamada & Tohkura, 1990). NJ listeners have also been shown to be sensitive to closure duration and transition duration when perceiving /r–l/ (Aoyama et al., 2004; Hattori & Iverson, 2009; Underbakke et al., 1988). Importantly, NJ listeners’ accurate identification of natural speech /r–l/ is best predicted by their use of the F3 cue (Gordon et al., 2001; Hattori & Iverson, 2009), suggesting that greater reliance on this cue might result in more NE-like performance.

Iverson and colleagues (Iverson et al., 2005) sought to correct this disparity in perceptual cue weightings. Their stimuli were based on the high variability natural speech stimuli described above (Bradlow et al., 1997; Bradlow et al., 1999; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991). They manipulated these tokens to increase the salience of F3 onset frequency. Despite these efforts to emphasize F3 onsets, the results were very similar to those found in earlier work: individuals were better able to identify /r–l/ tokens by all talkers, but not at the levels of NE listeners. They also found no changes in F3 sensitivity following training. This may be due to the continued presence of cues in the training stimuli that may be more salient to NJ listeners – cues that are partially consistent with but not the most reliable indicators of /r–l/ category membership (Lotto et al., 2004; O'Connor et al., 1957; Yamada & Tohkura, 1990).

It is apparent that though F3 onset frequency is the most reliable cue to /r–l/ category membership, NJ listeners have difficulty relying on this cue to differentiate /r–l/ even after training (Iverson et al., 2005) or after immersion in English (Aoyama et al., 2004; Gordon et al., 2001). It may be that the presence of cues other than F3 (e.g., F1 and F2) in training stimuli and in natural speech allows NJ listeners to rely on less reliable cues. In this article we consider the possibility that the presence of non-F3 cues in other training studies (and in natural immersion) might allow NJ speakers to rely on cues other than F3. From that perspective, we examine whether focusing all variation amongst /r–l/ tokens on F3 onset frequency will allow NJ speakers to learn to rely on F3. Specifically, if all remaining cues are equivalent among training stimuli, leaving only F3 onset frequency as a cue to differentiate among instances of /r/ and /l/, this may lead NJ learners to learn to rely on this cue, and this in turn should result in NE-like identification and discrimination performance. Our hope was that we might then see both the high levels of improvement on trained stimuli and generalization to untrained natural speech /r–l/ that have eluded previous efforts.

Experiments 1a and 1b

We constructed four synthetic /r–l/ series differing only on vowel: /ra–la/, /ræ–læ/, /ri–li/ and /ru–lu/, training participants with stimuli drawn from the /ra–la/ series and testing all participants with all four vowel contexts to assess generalization.[1] Within each series, only the F3 onset frequency varied; all other formants and transition durations were held constant.

We relied on a training procedure similar to that used in the condition of McCandliss et al. (2002) in which NJ participants exhibited the greatest improvement in English /r–l/ perception. This condition relied simply on repeated presentations of two fixed, moderately difficult stimuli with feedback. Based on the native-listener identification curves shown in Figure 2 below, we adopted Stimulus 4 and Stimulus 12 (circled in the figure) as moderately difficult /r/ and /l/ stimuli in Experiment 1a. As we shall see, participants in Experiment 1a did not improve from pre- to post-test. Therefore, in Experiment 1b, we used Stimulus 1 and Stimulus 16 as training stimuli. Since the results were similar in these two experiments, we present them together in the following section.

In addition to testing for improvements in identification for trained and untrained vowel contexts and natural speech, we looked for a shift to more NE-like discrimination post-training. This would be marked by an increase in discrimination accuracy at the /r–l/ category boundary (series middle) relative to within-category discrimination (series end).

Method

Participants

Sixteen NJ volunteers were recruited from the Pittsburgh area and participated in return for payment (see also footnote 1). As described in Footnote 1, data from those four individuals trained on /ri–li/ are not reported here. All reported normal hearing. There is no information regarding musical ability or years English was studied.

Participant eligibility was judged by performance in an English /r–l/ discrimination pre-test. Those participants scoring greater than 70% correct were excluded from participating in the remainder of the experiment (McCandliss et al., 2002). Four participants were excluded on this basis, noting that no participant performed at NE-like levels (ceiling). The first four eligible participants were used in Experiment 1a; the next four were used in 1b. Comparisons of eligible versus ineligible participants on age (31.75 vs. 30.25 years, $t(9) = 0.47, p = .65$), length of residency in North America (2.48 vs. 1.56 years, $t(9) = 0.77, p = .46$), age of first learning English (12.62 vs. 12.5 years old, $t(6) = 0.36, p = .73$), and self-reported ratios of spoken English to spoken Japanese (1.81 vs. 0.84 English/Japanese ratio, $t(9) = 0.59, p = .57$) revealed no across-group differences. Within each experiment, participants were divided equally into trained or untrained groups; group assignments were random.

Materials

Synthesized speech stimuli

Four 16-step synthesized consonant–vowel (CV) speech series varying from English /r/ to /l/ were created. The series were distinguished by the vowel, /a/, /æ/, /i/, and /u/. Within a series, only the third formant (F3) onset frequency distinguished members of the series. Stimuli were sampled at 11025 Hz and RMS matched in amplitude.

Syllables were synthesized using the parallel branch of the Klatt synthesizer (Klatt, 1980; Klatt & Klatt, 1990). Each stimulus was 330 ms in total duration, with silence for the first 10 and last 5 ms. The fundamental frequency

[1] Four additional participants (two in Experiment 1a and two in Experiment 1b) were trained using stimuli from the /ri–li/ continuum. It became apparent during the course of the experiment that the /ri–li/ stimuli were especially difficult for NJ participants. The high F2 frequency in the /ri–li/ training stimuli may interfere with NJ listeners' ability to access the F3 cue (Travis Wade, personal communication). Because of this, we used only /ra–la/ in training in our second experiment. Since Experiment 1 produced no training effect for training with either /ra–la/ or /ri–li/, the main motivation for reporting it is to compare the results of Experiment 1 to Experiment 2. Therefore, we report only the results of training with /ra–la/ in Experiment 1.

(*f*0) was a constant 110 Hz. The first and second formants (F1 and F2) had onset frequencies of 478 and 1088 Hz, respectively, and held these values across 85 ms at which time they linearly transitioned to the vowel steady-state frequency across 95 ms. F1 amplitude transitioned linearly from 0 to 50 dB across 35 ms whereas F2 amplitude transitioned linearly from 0 to 55 dB across 70 ms.[2] The fourth formant (F4) had a steady-state value of 3850 Hz across the duration of the sound. The amplitude of F4 transitioned linearly from 0 to 20 dB across 125 ms.

Within series, stimuli were distinguished by the F3 onset frequency, which varied from 1601 to 3400 Hz in increments of 43 Mel steps. F3 was steady-state at these values for 65 ms, then linearly transitioned to 2530 Hz across 115 ms. It remained at this frequency for the duration of the stimulus. F3 amplitude at stimulus onset covaried with onset frequency, varying from 60 (at 3400 Hz) to 45 dB (at 1601 Hz) in 1 dB steps. F3 onset amplitude began linearly transitioning to the vowel steady state (60 dB) at 65 ms and reached 60 dB at 180 ms.

The four /r–l/ CV series were distinguished by the final vowel. The vowels /a/ and /æ/ shared an F1 frequency of 705 Hz whereas /i/ and /u/ shared an F1 frequency of 205 Hz. F2 frequency for /i/ and /æ/ was 2005 Hz. The vowels /a/ and /u/ shared a F2 steady-state frequency of 1035 Hz, but /u/ F2 began at 1450 Hz (180–210 ms) before linearly transitioning over the next 50 ms to the steady-state value.[3] All steady-states and transitions had identical durations both across and within CV series, removing duration as a possible cue to category membership (Aoyama et al., 2004; Iverson et al., 2005). Thus, /a/ and /æ/ vowel contexts differed from one another along the F2 dimension whereas /æ/ and /i/ differed from one another along the F1 dimension. This orthogonality provided the opportunity to examine the effects of each dimension on generalization. Note that /a/ and /i/ differ from one another along both dimensions (as do /æ/ and /u/). Pseudo-spectrograms of the synthesis parameters can be found in Figure 1.[4]

To assure that these synthesized stimuli were reliably labeled as /r/ and /l/, 13 NE monolingual listeners responded to 15 repetitions of each of the 64 stimuli (4 series × 16 stimuli) as “r” or “l”, presented in random order mixed across vowel context. Identification curves are shown in Figure 2, demonstrating reliable, if imperfect, identifications as /r/ and /l/ by English listeners for these stylized synthetic speech stimuli. The imperfect identifications at the /l/ end of the series may be due in part to the high F3 onset frequencies for /l/ (3400 Hz at the most extreme). This value is within the range found in natural productions (Lotto et al., 2004) but is higher relative to the vowel than is typical (O’Connor et al., 1957). However, an examination of the data at the individual level indicates that most of the listeners reliably divided the stimuli into /r/ and /l/ categories and what appears to be imperfect categorization at the /l/-end of the series is driven by two listeners who identified most of the stimuli as /r/.

We used the native English listeners’ identification curves’ to identify their /ra–la/ category boundary (the /ra–la/ curves being the steepest). A proportion /r/ response difference of .50 or greater between two members of a discrimination pair was indicative of a category boundary and called the SERIES MIDDLE (stimuli pairings 6–10 and 7–11). The remaining pairings were classified as the SERIES END. Position assignment from the /ra–la/ series was extrapolated to the other vowel contexts. These classifications were used when analyzing the discrimination tests to determine if NJ listeners showed better between-category than within-category discrimination, as would be expected by native English listeners (Liberman, Harris, Hoffman & Griffith, 1957).

Natural speech stimuli.

Two lists of 16 /r–l/ English minimal pair words were created, resulting in 32 total pairs. Two native English speakers, one male and one female, produced all words in each list, for a total of four speakers. One male speaker was fluent in German, which he began learning at age 18, with English his only language prior to this. The remaining three speakers identified themselves as monolingual English speakers. The full list of minimal pairs is shown in Table 1. It is divided into four pair types, based on Logan et al. (1991), based on /r–l/ position: initial singleton (*lock–rock*; 9 pairs), initial cluster (*flesh–fresh*; 8 pairs), intervocalic (*elect–erect*; 7 pairs), and final singleton (*file–fire*; 8 pairs). Within each list of 16 words, each member of a minimal pair was spoken by a different talker; all 32 words were presented at test.

Talkers produced two exemplars of each word in the sentence, “The next word is _____, _____”. Words were recorded at 11025 Hz on a PC desktop running Windows XP. The second production of each word was chosen as the stimulus. Stimuli were RMS matched in amplitude.

Procedure

McCandliss et al. (2002) found the greatest improvement when training consisted of repeated presentations of two moderately difficult stimuli combined with performance

[2] The exception to this was the amplitude of F2 in /ri–li/ context. Spectral analyses revealed that the standard synthesis parameters produced higher-amplitude F2 in this context. Therefore, for this context, F2 amplitude transitioned from 0 to 45 dB instead of 0 to 55 dB. This manipulation of synthesis parameters produced acoustically more similar stimuli across vowel series.

[3] This acoustic manipulation was deemed necessary to more closely mimic natural consonants and to produce reliable /ru–lu/ percepts among English listeners.

[4] An additional stimulus for each vowel context was created for the identification test of Experiment 1b. This stimulus had an F3 onset frequency of 1514 Hz and an onset amplitude of 61 dB.

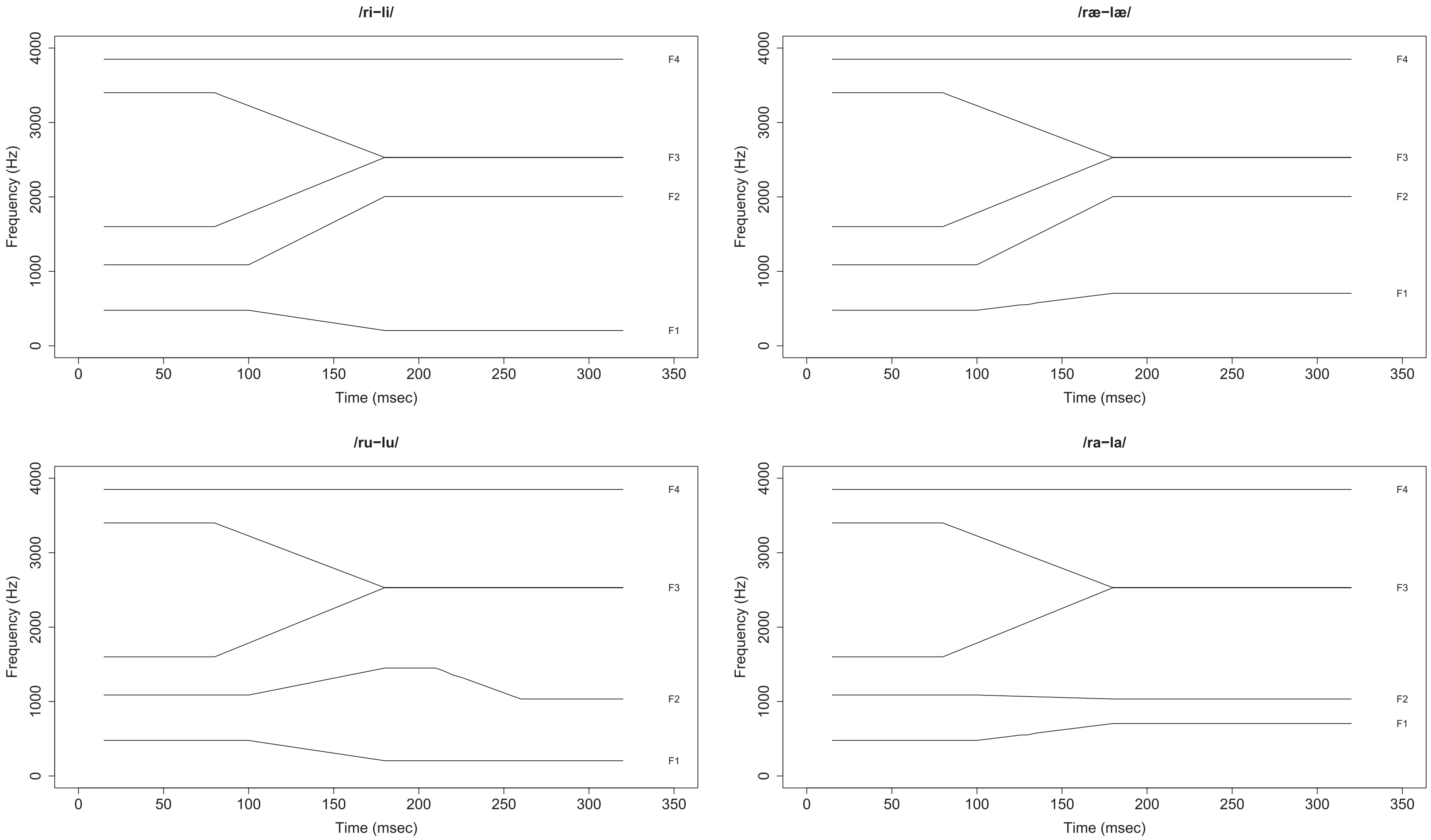

Figure 1. Endpoints of the /r–l/ series for each of the four (/a/, /æ/, /i/, and /u/) vowel contexts representing onsets and trajectories of the first through fourth formants. Stimuli within a vowel context differed from one another in only F3 onset frequency.

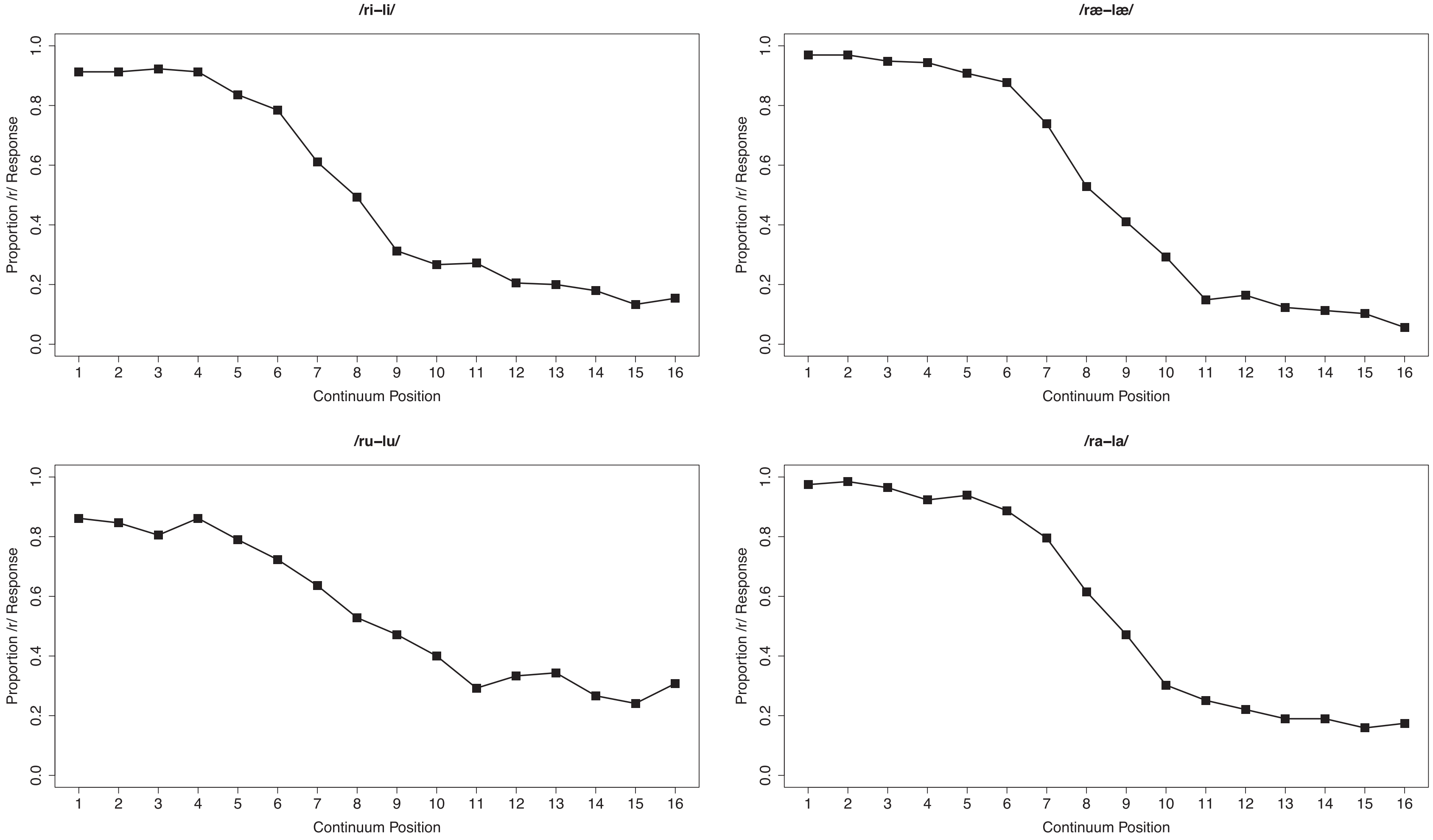

Figure 2. Proportion /r/ response for each stimulus by native English (NE) listeners. Stimuli are presented here separated by vowel context; participants heard all stimuli intermixed. Those stimuli that were used for training in Experiment 1a are marked with large circles.

Table 1. *Minimal pair words to test participants' perception of natural speech, divided by the position of the /r–l/ contrast.*

Word initial	Consonant cluster	Intervocalic	Word final
rack – lack	breed – bleed	aright – alight	bare – bale
raw – law	broom – bloom	arrive – alive	dare – dale
red – led	crash – clash	array – allay	fire – file
rice – lice	crowd – cloud	arouse – allows	hear – heal
rid – lid	fresh – flesh	berated – belated	mire – mile
road – load	fright – flight	erect – elect	peer – peel
rock – lock	grow – glow	pirate – pilot	steer – steel
room – loom	pray – play		tire – tile
rust – lust			

feedback. This procedure served as the basis for Experiment 1 training, using stimuli in which instances of /r–l/ differed only on F3 onset frequency. Based on the native-listener identification curves shown in Figure 2, we adopted Stimuli 4 and 12 as moderately difficult /r/ and /l/ stimuli, respectively (Experiment 1a, N = 4). These stimuli are marked with circles in Figure 2. Following these listeners' failure to learn, in a subsequent effort (Experiment 1b; N = 4), we used Stimuli 1 and 16 as training stimuli.

NJ participants began with an eligibility AX discrimination test using the training stimuli (Stimulus 4 versus Stimulus 12 in Experiment 1a; Stimulus 1 versus Stimulus 16 in Experiment 1b) across 25 repetitions. Individuals with discrimination accuracy greater than 70% were deemed ineligible for training (N = 4 ineligible individuals, accuracy *M* = 74%).

Eligible participants (N = 8, accuracy *M* = 51%) next completed an identification test of the synthesized speech stimuli, followed by a discrimination test of the same stimuli, then a test of natural speech identification. All participants would repeat these three tests on Day 11, the final post-test day. Additionally, on Day 6 all participants repeated the identification test of synthesized speech and the discrimination test of synthesized speech. The intervening days, Day 2–Day 5 and Day 7–Day 10, were the training days. Those individuals assigned to a training condition received a laptop computer and trained at home. Details of these testing and training procedures are described below. Logs of time-on-task were monitored to assure participants' compliance with the training program.

All tests were administered under the control of E-Prime (Psychological Software Tools, Pittsburgh PA) on a PC laptop running Windows XP. Stimuli were presented diotically over Beyer DT-150 headphones at approximately 70 dB.

Identification

The identification test included all four synthesized /r–l/ series and was blocked by vowel context; no feedback was given. Participants in Experiment 1a heard 12 presentations of Stimuli 4, 6, 8, 10, and 12 (bounded by training stimuli 4 and 12) along each series and indicated whether the syllable began with /r/ or /l/ by pressing a labeled key on the keyboard. Short breaks were provided between blocks. Participants in Experiment 1b completed the same test for even-numbered series members (Stimuli 0–16; Stimulus 0 was an additional stimulus created to equalize the number of presented stimuli on either side of the series midpoint, Footnote 3). All participants took the identification test on Day 1 (pre-test), Day 6, and Day 11 (post-test).

Discrimination

Immediately following the identification test, an AX discrimination task of the synthesized speech series was administered. On each trial, listeners heard a pair of stimuli made up of items 4 steps apart along the series (e.g. stimuli 4 and 8 were used to create two same pairs: 4–4 and 8–8; and two different pairs: 4–8; 8–4). Listeners heard four presentations of each pair; pair members were separated by a 750 ms ISI. Stimuli 3–13 from all CV series were used in the discrimination test in Experiment 1a; Stimuli 1–16 were used in Experiment 1b. Participants indicated if the pair was "Same" by pressing "1" and "Different" by pressing "0" on the keyboard. Tests were blocked by vowel. Short breaks were provided between blocks. All participants took the discrimination test on Day 1, Day 6, and Day 11.

Natural speech identification

Identification of the natural speech English /r–l/ minimal pairs was assessed at pre- and post-test; identifications were made without feedback. On a given trial, orthographic representations of each word were presented on the computer monitor while the acoustic stimulus was presented once. Listeners used the keyboard to indicate which pair member was presented. Stimuli were not repeated within a test and each member of a minimal pair was spoken by a different talker.

Training

Individuals participating in training conditions completed daily 30-minute sessions at home using PC laptops running Windows XP and Beyer DT-100 headphones provided to them. Participants were given a sheet explaining the details of the daily training regimen in both English and Japanese and each training session was checked for completeness at the mid-test and post-test.

NJ participants trained with /ra–la/ synthesized speech. The training task was a 2AFC identification task. On each trial, listeners heard one of the two training stimuli (Stimulus 4 or 12 in Experiment 1a; Stimulus 1 or 16 in Experiment 1b, differentiated exclusively by F3 onset frequency) and indicated whether it sounded like “r” or “l” by pressing a labeled button on the keyboard. Sounds were not repeated within a trial and training did not advance until a response had been made. Response accuracy feedback was visually presented for 500 ms following each response. In each training session, participants heard 10 repetitions of each training stimulus (Stimuli 4 and 12 in Experiment 1a; Stimuli 1 and 16 in Experiment 1b) in 25 blocks for a total of 500 training trials per day (McCandliss et al., 2002). Each training session took approximately 30 minutes to complete, with short breaks given between blocks.

Participants trained across eight days, not training on the days they took the mid-test and post-test. Thus, they attempted to categorize to 4,000 repetitions of the two training stimuli (2,000 repetitions of each stimulus), with accuracy feedback on each trial.

Results

Preliminary analyses revealed no post-test differences between Experiments 1a and 1b. Therefore, results are collapsed across these variables in all subsequent analyses. To assess the impact of training, we submitted the data from the identification, discrimination, and natural speech tests to a Training (Trained vs. Untrained) × Session (pre- vs. post-test) mixed model ANOVA, focusing attention on the crucial Training × Session interaction. Considering first the identification of synthetic speech, there was no Training × Session interaction, $F(1,10) = 0.13$, $p = .72$. Similarly, considering the discrimination data, there was no Training × Position (series end vs. series middle) × Session interaction, $F(1,10) = 0.01$, $p = .94$. We also found no Training × Session interaction for the identification of natural speech, $F(1,9) = 0.45$, $p = .52$. Thus, it appears that participants failed to improve in their ability to identify or discriminate /r–l/ tokens even within the trained vowel context; they also failed to improve in their ability to identify natural /r–l/ stimuli. This failure to learn is made more remarkable by the fact that (1) training tokens differed from one another in only F3 onset frequency, presumably enhancing the salience of this cue; (2) listeners heard 4,000 training tokens which they identified with feedback; (3) previous studies using a single continuum of training stimuli have repeatedly found evidence of improvement on the trained task and context even when no generalization was found (McCandliss et al., 2002; Strange & Dittman, 1984).

Despite the lack of evidence of a training effect at the group level, we found extensive individual differences in performance, consistent with earlier work in second-language acquisition and perceptual learning (Ingvalson & Wenger, 2005; Iverson, Ekanayake, Hamann, Sennema & Evans, 2008; Jenkins, Strange & Polka, 1995; Maddox, Diehl & Molis, 2001; Romaine, 2003). It was therefore possible that some individuals had learned to differentiate /r–l/ but their performance was masked by the group variability. We consequently undertook analyses at the level of the individual (see Supplementary Materials to be found on the journal’s webpage along the online version of the present paper). We did observe some differences from pre- to post-test among a number of participants, but such changes were limited. Since changes occurred for both Trained and Untrained individuals, these analyses provided no clear evidence that any participants benefited from training.

Discussion

Recent studies have pointed to the importance of F3 onset frequency as a reliable acoustic cue to /r–l/ category membership in NE speakers’ productions (Lotto et al., 2004) and perceptual identifications (Iverson et al., 2003; Yamada & Tohkura, 1990). NJ individuals speaking or perceiving English make much less use of this cue, leading to the hypothesis that shifts in perceptual cue weighting toward F3 may facilitate /r–l/ perception (Holt & Lotto, 2006; Iverson et al., 2005). Previous studies investigating laboratory-based training of NJ listeners on /r–l/ have typically employed stimuli that vary along multiple acoustic dimensions (Bradlow et al., 1997; Iverson et al., 2005; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991; McCandliss et al., 2002; Strange & Dittman, 1984). The presence of such acoustic variability prevents assessment of listeners’ usage of the F3 onset cue, and suggests that non-NE-like performance may be the result of learning to use acoustic cues other than F3 onset frequency. Here, we investigated whether training explicitly on F3 onset frequency alone would allow NJ participants to better categorize English /r–l/.

Training stimuli were differentiated only on F3 onset frequency and NJ listeners heard 2,000 instances of each stimulus distributed across eight days of training. For each instance, listeners were provided with immediate accuracy feedback. Yet, participants showed no evidence of an influence of training from pre- to post-test identification, discrimination, and natural speech identification. Although limited differences from pre- to post-test were observed in the data of a number of participants, such changes occurred for individuals in both Trained and Untrained groups. This lack of a training effect serves to highlight the considerable difficulty NJ listeners have in using F3 onset frequency to perceive English /r–l/ (Miyawaki et al., 1975).

Experiment 2

It is possible that the presence of other cues (e.g., closure duration and F2), even if they are held constant between instances of /r/ and l/, somehow prevents NJ listeners from being able to attend to the F3 cue (Iverson et al., 2005; Takagi, 2002; Takagi & Mann, 1995). In support of this possibility, it is interesting to note that NJ listeners are able to discriminate single, isolated F3 formants like those from minimal /ra–la/ syllables (Miyawaki et al., 1975). These non-speech sounds lack the harmonic structure of lower and higher frequencies to cause them to be heard as speech and are instead heard as non-speech beeps. These data demonstrate that NJ listeners' difficulty in using F3 onset as a cue to the English /r–l/ contrast does not arise from complete insensitivity to F3 onset frequency. More generally, it is possible that the training paradigm employed in Experiment 1 did not optimally promote learning. Even though Experiment 1b used series endpoints, it may be that these stimuli were not sufficiently distinct for our NJ participants to use as a starting point for learning. Our aim in Experiment 2 was to determine if starting with F3-only stimuli that listeners could differentiate and gradually restoring the remaining formants as performance improved could effect learning to use F3 in the /r–l context; McCandliss et al. (2002) found improvement using adaptive training both with and without the presence of feedback, leading us to also manipulate this variable. Comparing the results of this experiment to the total lack of improvement in Experiment 1 – which used full-spectrum /r–l/ throughout – would reveal whether F3-only training was beneficial relative to F1–F4 conjunction training where only F3 varied.

Method

Participants

Twenty-six NJ individuals from the Pittsburgh area, all of whom reported having normal hearing, participated in exchange for payment. There is no information regarding musical experience or years of English study. Ten participants (length of residency $M = 1.38$ years; age at test $M = 28$ years) were ineligible to participate in training because they were able to discriminate at least one set of the training stimuli from Experiment 1b at an accuracy level of greater than 70% ($M = 73\%$); again, no participant performed at NE-like levels. Comparisons of eligible versus ineligible NJ speakers on age (30.56 vs. 27.70 years, $t(16) = 0.95$, $p = .36$), length of residency in North America (1.62 vs. 1.38 years, $t(15) = 0.44$, $p = .67$), age of first learning English (12.0 vs. 12.4 years old, $t(20) = 0.76$, $p = .45$), and self-reported ratios of spoken English to spoken Japanese (3.31 vs. 0.81 English/Japanese ratio, $t(26) = 1.93$, $p = .06$) revealed no significant differences.

Of the remaining 16 participants (length of residency $M = 1.62$ years; age at test $M = 31$ years), four were assigned to the untrained group. The testing conditions for Experiment 2 were identical to those of Experiment 1b, so the two Untrained participants from Experiment 1b were also included in the Untrained group for Experiment 2 (we did not include the Untrained participants from Experiment 1a since their testing stimuli excluded the extreme end-points).

The remaining 12 participants were equally divided into feedback and no-feedback groups; group assignment was random. Thus, at test, six participants were trained on /ra–la/ with feedback, six participants were trained on /ra–la/ without feedback, and six individuals received no training. The number of participants per condition, while small, is within the range in earlier work, ranging from 6 trained listeners (Logan et al., 1993; Logan et al., 1994) to 19 trained listeners (Lively et al., 1994).

Materials

The training stimuli were based on the training stimuli in Experiment 1b and created using the parallel branch of the Klatt speech synthesizer (Klatt, 1980; Klatt & Klatt, 1990). Stimuli consisting of single-formant F3 trajectories were created from series endpoints (Stimuli 1 and 16). These were created by setting the amplitudes of all non-F3 formants to 0 dB, leaving one stimulus with an F3 trajectory that started at 1601 Hz and another that started at 3400 Hz. These stimuli thus possessed the same F3 as the full-spectrum speech stimuli of Experiment 1, but did not have acoustic energy in frequencies other than F3 and sounded like non-speech beeps.

To implement the adaptive training, twenty additional stimuli were created. For these stimuli, the amplitude of the speech spectrum that was damped to 0 dB in the single-formant F3 trajectories was gradually restored by increasing the amplitudes of F1, F2, and F4 in increments of 10% of their final values until they reached the values of Experiment 1 stimuli (making 10 increases in 10% steps of the amplitude values of Experiment 1 stimuli for each endpoint, resulting in 20 stimuli). Figure 3 shows

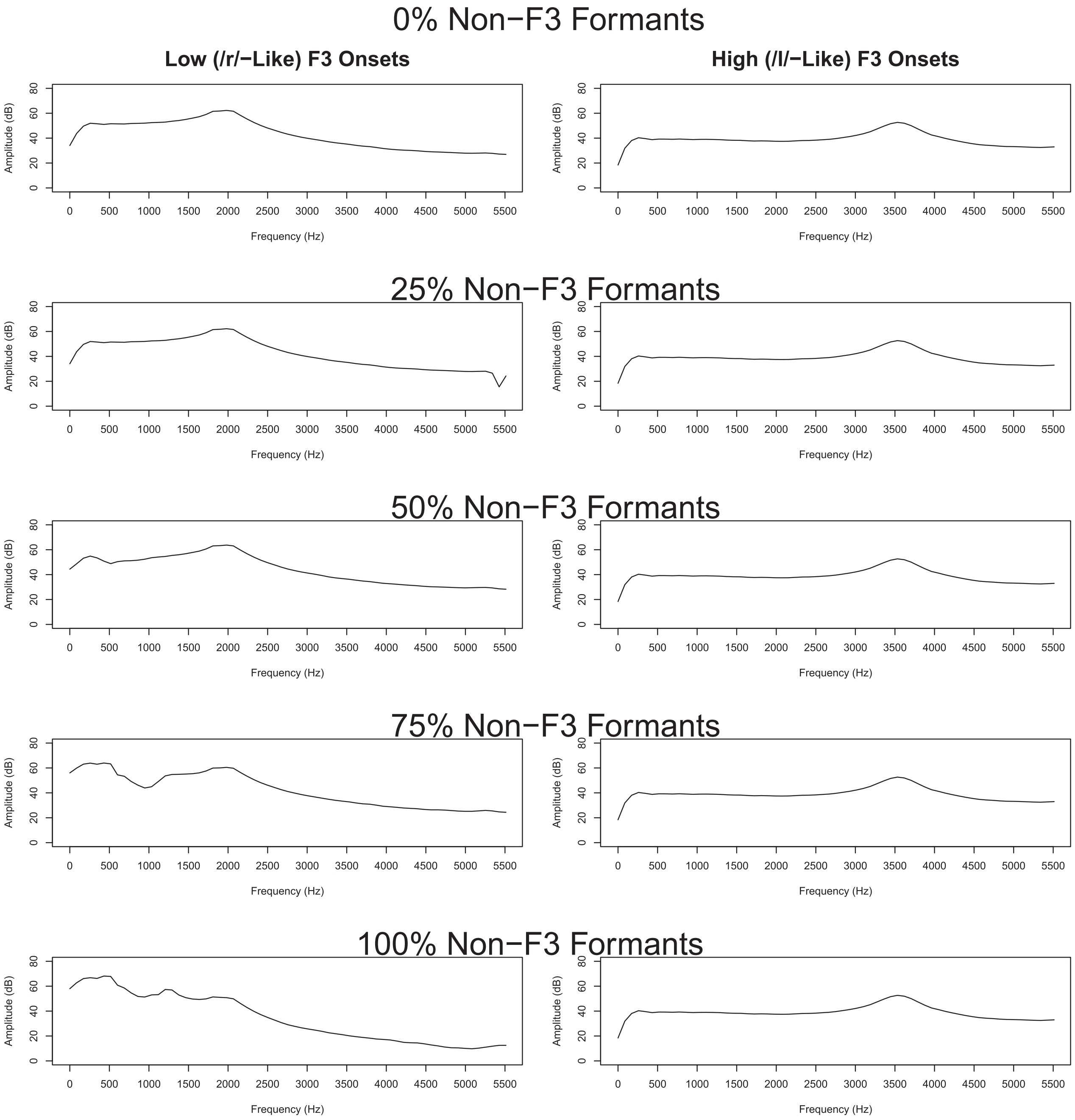

Figure 3. Frequency × amplitude plots of the /ra–la/ adaptive training stimuli from Experiment 2 when the other formants are at 0%, 25%, 50%, and 100% of their final values.

frequency × amplitude plots when the non-F3 formants are at 0%, 25%, 50%, and 100% of their final values. All modifications were limited to the endpoints of the series.

Stimuli were sampled at 11025 Hz and RMS matched in overall amplitude to Experiment 1 stimuli. Stimuli were presented to participants in the same manner as Experiment 1.

Procedure

Testing procedure and the number of training sessions and trials per session were identical to Experiment 1b.

Training

As in Experiment 1, each training trial consisted of an auditory stimulus, followed by a 2AFC response. In line with McCandliss et al.'s (2002) adaptive procedure, half the participants received feedback after each response and half did not receive feedback.

Training began with the presentation of single-formant F3 trajectories that modeled Stimuli 1 and 16 of Experiment 1. Participants were told to respond “r” to the most /r/-like stimulus (F3 onset below the vowel) and to respond “l” to the most /l/-like stimulus (F3 onset above the vowel). Because these stimuli were non-speech and may have been difficult for participants to map onto the “r” and “l” response options, all participants, regardless of training condition (Feedback or No-Feedback) were given feedback when hearing these stimuli to ensure the response options were correctly mapped. From this point forward the stimuli were determined by the participants’ response accuracy. If the participant made six consecutive correct responses, the stimuli became progressively more speech-like. If the participant made one incorrect response, the stimuli became progressively less speech-like. Starting with single-formant F3-trajectories, stimuli were made more speech-like by increasing the amplitude of the non-F3 spectrum by 10%, edging the stimuli toward full-spectrum speech; stimuli were made less speech-like by decreasing the amplitude of non-F3 spectrum by 10%. If the participant made six correct responses to full-spectrum /r–l/, stimuli were then made progressively more difficult by moving closer to the center of the series (Figure 1 above). For example, the first full-spectrum stimuli would be the endpoints, Stimuli 1 and 16. Six consecutive correct responses to these stimuli would move the participant to Stimuli 2 and 15; the participant would then progress to stimuli 3 and 14 or 1 and 16, depending on performance. Overall, the adaptive procedure allowed the stimuli to range from pure isolated F3 formants to full spectrum stimuli in which the difference between F3 onsets of /r/ and /l/ tokens was reduced to only 55.05 Hz; Stimuli 7 vs. 9.

Trials were initiated with a visual fixation cross that preceded acoustic presentation of the sound. Participants identified the sound as “/r/-like” or “/l/-like”. No limits were placed on response time, but participants were encouraged not to deliberate at length. Stimuli were randomized with the constraint that low and high F3 onsets were presented equally often across the six trials required for advancement. Following the response, a blank screen appeared for 500 ms for participants in the No-Feedback condition. Those receiving feedback received an indicator of their accuracy for 500 ms. Each participant completed approximately 500 trials each day (with some variability because the task was adaptive) and began training in the next session with stimuli based on performance at the end of the preceding session.

Results

There are three main findings from this experiment. First, participants in both the Feedback and No-Feedback groups improved over the course of training, with considerable within and between subject variability. Second, three Feedback participants and one No-Feedback participant showed clear evidence of a pre- to post-test transition to NE-like identification on the trained stimulus series, while no Untrained participants showed such changes. Of the four participants showing a training effect, two showed evidence of an acquired discrimination peak at the /ra–la/ category boundary. There was evidence of generalization of the training effect to identification on the /ræ–læ/ series in only one of these participants, but three of the four showed evidence of an improvement from pre- to post-test in the natural speech identification test. Overall, training was a partial success for some participants. We now proceed to present these findings in detail, beginning with the effect of training.

Training

We assigned a difficulty score to each pair of training stimuli, using the percentage of the non-F3 formants present in the stimulus for the ten pairs in which this percentage varied (assuming that the inclusion of additional /r–l/ information would make the stimuli more difficult for NJ listeners), and then assigning difficulty scores incrementing in 10-point steps for the eight remaining pairs of stimuli, which stepped inward from the series endpoints. Thus, the maximum difficulty level was 180, assigned to the stimuli just one step apart the middle stimulus. The average difficulty level across different points in training is shown in Figure 4. Performance improved within sessions and tended to drop from the end of one session to the beginning of the next. The between-session drop shows signs of getting smaller in the later sessions, indicating maintenance of identification performance between sessions later in training.

We submitted the data to a Session (1–8) × Group (Feedback or No-Feedback) mixed-model ANOVA. We found a main effect of session, $F(7{,}27) = 5.17$, $p < .001$, $\eta^2 = 0.124$, indicating that the overall improvement is reliable across participants. Although the No-Feedback group appears to have reached higher difficulty levels on average over the last five sessions, neither the main effect of group nor the group by session interaction were significant. We supplemented the global ANOVA with an ANOVA for a linear trend, performed separately for each training group. Both the Feedback ($F(1{,}27) = 4.36$, $p < .05$, $\eta^2 = 0.076$) and No-Feedback ($F(1{,}27) = 25.57$, $p < .001$, $\eta^2 = 0.449$) groups significantly improved over the course of training.

Trained vowel series

Identification

Initial inspection of the pre- and post-test identification data revealed large individual differences among participants, obscuring possible training effects at the

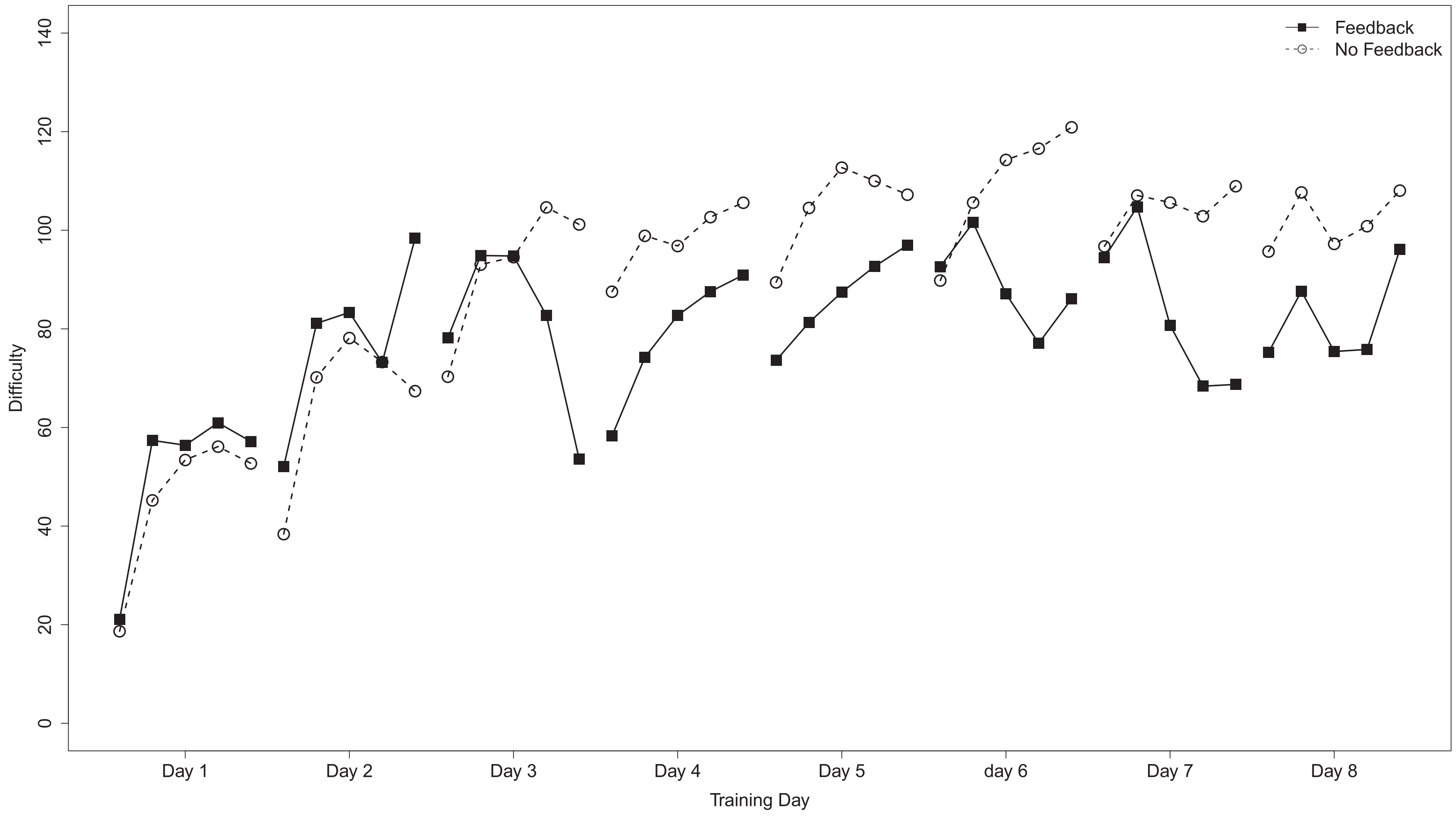

Figure 4. Level of difficulty for participants trained either with or without feedback as a function of training day. The y-axis is an indicator of difficulty: contexts where it is more difficult for native Japanese (NJ) listeners to differentiate high and low F3 onset frequencies are further from the origin. The portion of the axis marked “Amplitude” refers to those stimuli where non-F3 formants amplitudes are manipulated; values here indicate the percentage of the non-F3 formants present in the stimuli. The portion of the axis marked “Series” refers to those stimuli where non-F3 formants are fully present but the difference between low and high F3 onset frequencies has been reduced; values here indicate the difference (in Hz) between the onset frequencies.

individual level. Furthermore, we often found large changes from pre to post-test that differed from the expected pattern of change from a relatively flat pre-test identification function to a relatively steep sigmoidal identification function. We therefore considered the data of each participant separately, looking for evidence of learning at the level of the individual, employing three separate criteria. Only participants meeting all three criteria were treated as showing a true shift from initial insensitivity to F3 to NE-like sensitivity.

The first criterion relied on the chi-square statistic to determine whether there was a reliable difference between pre- and post-test, combining the evidence of a pre- vs. post-test difference at each point along the 16-step stimulus series into a single value of chi square (Agresti, 1992).[5] The second criterion relied on the logistic regression slope to assess whether the change from pre- to post-test could be characterized as an increase in the slope of the identification function. A single z-score quantified the difference between the pre- and post-test slopes for each individual listener:

$$\frac{(Slope_{post-test} - Slope_{pre-test})}{\sqrt{(SE^2_{post-test} + SE^2_{pre-test})}}$$

Comparing these z-scores to the normal distribution provides a significance test for the difference between these slopes. Chi-square test results, pre- and post-test slopes, standard errors, and the resulting z-scores for all listeners on all series can be found in Table S.5. Listeners' identification curves at the pre- and post-test can be found in Figure S.1 (in supplementary materials), which shows both the extensive variability amongst listeners and the myriad identification patterns seen. Using just these two criteria, three Feedback listeners – listeners Feedback-1 ($\chi^2 = 63.13$, $p < .001$; $z = –2.86$, $p < .05$), Feedback-3 ($\chi^2 = 43.13$, $p < .001$; $z = –4.94$, $p < .05$), and Feedback-5 ($\chi^2 = 30.96$, $p < .05$; $z = –3.58$, $p < .05$) – and two No-Feedback listeners – listeners No-Feedback-4 ($\chi^2 = 32.86$, $p < .05$; $z = –3.77$, $p < .05$) and No-Feedback-6 ($\chi^2 = 52.70$, $p < .001$; $z = –3.08$, $p < .05$) – all showed an effect of training identifying stimuli in the /ra–la/ series.

Our final criterion is needed to address the pattern seen in participant No-Feedback-6 (Figure S.1). To visual inspection, both the pre- and post-test results show a steep transition in identification performance across the /ra–la/ series. The pre-test logistic regression slope is less extreme than the post-test value, but this is an artifact of the uptick in /r/ responses at the extreme /l/-like end of the series.[6] While such changes may well reflect an influence of training, it is not clear that in this case there has been a change in the participant's sensitivity to differences among stimuli along the /ra–la/ series. To allow separation of such cases, we adopted a final subjective filter to determine if participants who met the first two criteria showed a discernable increase in identification function slope from pre- to post-test. Visual inspection of the other four cases passing the first two criteria – participants Feedback-1, Feedback-3, Feedback-5, and No-Feedback-4 – shows that the canonical pattern is observed in all of these cases. In all four cases, pre-test identification performance is relatively uniform across all stimuli, while post-test performance shows the canonical NE-like pattern.

It is important to note that none of the Untrained participants showed changes that met both the first two criteria. One participant, Untrained-4, did show a significant change in the logistic regression slope ($\chi^2 = 21.54$, $p > .05$; $z = –3.49$, $p < .05$), meeting one of the criteria. However, this participant's data showed a similarly steep transition in the pre-test, with an uptick in the proportion or /r/ responses at the most extreme /l/-like end of the series. Thus, there is no evidence that any of the Untrained participants showed a significant increase in sensitivity to F3 from pre- to post-test.

Discrimination

We used two criteria to determine whether each participant learned to discriminate across the NE listeners' /r–l/ category boundary on the trained stimuli: (a) we compared discrimination performance on the middle of the /ra–la/ series at pre- and post-test; and (b) we asked whether post-test discrimination was better at the middle of the series – where a category boundary is spanned – than at the end of the series – where the tokens are within-category. Using paired t-tests to compare pre- and post-test performance on midpoint stimuli (Table S.6), several listeners showed significant improvement: Feedback-2 ($t(31) = –2.34$, $p = 0.02$, $d = 0.461$, pre-test accuracy = 53%, post-test accuracy 75%), Feedback-3 ($t(31) = –2.74$, $p = .03$, $d = 0.577$, pre-test accuracy 38%, post-test accuracy 66%), Feedback-5 ($t(31) = –2.74$, $p = .01$, $d = 0.577$, pre-test accuracy 50%, post-test accuracy 81%), No-Feedback-4 ($t(31) = –4.98$, $p < .001$, $d = 1.317$, pre-test accuracy 47%, post-test accuracy 97%), No-Feedback-5 ($t(31) = –2.74$, $p = .01$, $d = 0.649$, pre-test accuracy 38%, post-test accuracy 69%), and Untrained-2 ($t(31)$

[5] Because the expected numbers of correct or incorrect responses were often less than 5, Fisher's exact test was used to obtain a p-value at each continuum point. This was then converted to the corresponding normal deviate (z-score). These were squared and then summed across the continuum to obtain the overall Chi Square.

[6] The logistic regression slope variable is highly sensitive to a change in performance on one continuum endpoint stimulus. Such changes did occur in several participants (Feedback-2, No-Feedback-5, and Untrained-4). Some other participants showed an overall change from pre-to-post-test but not a change in slope (Feedback-4, Feedback-6, and No-Feedback-2). For these participants, neither pre- nor post-test performance was similar to that of NE participants.

= –2.18, p = .02, d = 0.585, pre-test accuracy 44%, post-test accuracy 72%). Comparing middle vs. end-point stimuli (Table S.7) revealed listeners Feedback-5 (t(190) = –1.86, p = .03, d = 0.383 middle accuracy 81%, end accuracy 64%), No-Feedback-4 (t(190) = –4.55, p < .001, d = 1.088, middle accuracy 97%, end accuracy 56%), and Untrained-6 (t(190) = –1.81, p = .04, d = 0.356, middle accuracy 66%, end accuracy 48%) all showed better discrimination of middle than endpoint stimuli.

Combining the two tests, two of the four participants who showed clear signs of a transition from relative insensitivity to NE-like identification performance – Feedback-5 and No-Feedback-4 – showed the expected pattern of discrimination performance on both measures. No untrained participants showed this pattern.

Untrained vowel series

Identification

Using the first two criteria for a training effect on identification performance on the untrained vowel continua, listeners Feedback-1 (χ^2 = 37.33, p < .01; z = –2.79, p < .05), No-Feedback-1 (χ^2 = 108.93, p < .001; z = –4.99, p < .05), No-Feedback-5 (χ^2 = 79.34, p < .001; z = –4.39, p < .05), and Untrained-2 (χ^2 = 32.78, p < .05, z = –3.45, p < .05) showed a significant change from pre- to post-test on the /ræ–læ/ series. Inspection of the pre- and post-test curves reveals that two of the participants (Feedback-1 and No-Feedback-5) met the third criterion, showing initially flat identification functions that became more NE-like with training. For these participants, then, we have some evidence that the benefit of training generalized from the trained /ra–la/ series to the untrained /ræ–læ/ series. Participant Untrained-2 showed some evidence of improvement from pre- to post-test, but the final pattern is quite weak in comparison to the changes seen in participants Feedback-1 and No-Feedback-5. Overall, the data hint at the possibility that some trained participants were able to generalize an effect of training on the /ra–la/ series to produce NE-like identification performance on the untrained /ræ–læ/ series (Figure S.2). There was little if any sign of improvement on any of the other series. Several participants showed significant pre- to post-test differences on the /ru–lu/ series; however, none of these listeners showed NE-like identification curves at the post-test (Figures S.3 and S.4). None of the participants who showed a significant improvement identifying the trained vowel context also showed improvement in identification with either the /ru–lu/ or /ri–li/ series. Thus, there is no basis for thinking the training effect generalized to these series. Because only two participants showed any evidence of a generalization effect, we hesitate to speculate why generalization may have occurred in the /ræ–læ/ context but not in the /ri–li/ or /ru–lu/ contexts.

Discrimination

When assessing discrimination performance on the untrained vowel contexts via paired t-tests, no trained listener showed a significant improvement across the NE boundary from pre- to post-test together with evidence of NE-like discrimination (greater between than within categories) at the post-test.

Identification of natural speech

As in Experiment 1, we assessed differences in performance on natural speech identification collapsing across /r–l/ position. Each listener's performance as a function of /r–l/ position can be seen in Table S.8.

Two of the four participants who showed a transition to more NE-like identification on the trained /ra–la/ series, Feedback-1 and Feedback-5, also showed a significant improvement from the pre- to the post-test (Feedback-1: t(63) = 2.81, p = .003, d = 0.390, pre-test accuracy 66%, post-test accuracy 82%); Feedback-5: t(63) = 2.68, p = .005, d = 0.416, pre-test accuracy 62%, post-test accuracy 81%). No other participants showed such clear evidence of improvement from pre- to post-test, although two participants, No-Feedback-5 and Untrained-3, showed some evidence of improvement (No-Feedback-5: t(63) = 2.01, p = .02, d = 0.307, pre-test accuracy 63%, post-test accuracy 77%; Untrained-3: t(63) = 1.72, p = .04, d = 0.276 pre-test accuracy = 76%, post-test accuracy 87%).

Summary

Positive evidence of a pre- to post-test improvement is largely restricted to the participants who were trained with feedback. Three of these six participants – Feedback-1, Feeback-3, and Feedback-5 – showed a transition to NE-like performance on the trained /ra–la/ series, and Feedback-1 and Feedback-5 showed further signs of improvement. Feedback-1 showed generalization of the NE-like transition to another synthetic speech series, /ræ–læ/, and Feedback-5 showed a transition to a NE-like pattern of discrimination on the trained /ra–la/ series. These two participants also showed robust evidence of a post-test improvement in identification of /r–l/ from natural speech.

Training without feedback produced unequivocal evidence of a transition to native-like identification performance in only one of the six participants, No-Feedback-4. This participant also showed evidence of a transition to NE-like discrimination on the trained series, but did not show generalization to untrained stimuli. One other participant in this group – No-Feedback-5 – showed a transition to a NE-like pattern of identification on the untrained /ræ–læ/ series, along with a statistically borderline improvement in natural speech identification. Inspection of this participant's pre- and post-test identification curves for the trained series shows some evidence of a training effect there as well, but it is not clear-cut. Thus, it appears that three participants trained with feedback and one or two participants trained without feedback showed a benefit of training; of these participants, two of those trained with feedback showed fairly clear evidence of an improvement in natural speech identification. As with any experiment utilizing repeated stimulus presentations, it is possible that the apparent learning effects are not the result of training but are instead the result of repeated stimulus presentations mitigated by attentional and motivational factors. However, the fact that only one untrained participant – Untrained-4 – showed any of the positive signs of improvement from pre- to post-test, in the form of a borderline improvement on identification of natural speech stimuli, suggests to us that this is not the case. Thus, there is reason to believe that the pre- to post-test changes, where observed, are due to the training regime, and not simply to an improvement due to test practice.

We now consider the relationship between the efficacy of the training and the performance of the participants during training, shown in Table S.9. Two participants who showed steady progress over training (Feedback-1 and No-Feedback-4) also showed a transition to NE-like identification. However, several participants who showed good progress during training nevertheless failed to show a transition to native-like performance, whereas other participants whose performance during the training phase was quite variable nevertheless appeared to benefit from training. For example, participant Feedback-5 showed clear pre- to post-test improvement while showing quite variable performance during training; participant Feedback-3 also showed a transition to NE-like identification after variable performance during training. Both of these participants did have some sessions, however, during which their training task performance was near the upper end of the range, while none of the three participants whose performance during training stayed uniformly below difficulty level 100 showed a transition to NE-like performance.

Discussion

NJ listeners are able to discriminate high and low F3 onset frequencies outside of the /r–l/ context (Miyawaki et al., 1975), even when they cannot use F3 to distinguish /r–l/ in natural or synthetic speech. Using this, together with the idea that progressing from an easy to a hard differentiation can lead to successful learning (e.g., McCandliss et al., 2002), we have attempted to teach NJ speakers to use the F3 cue to differentiate /r–l/, starting with isolated F3 formants and then gradually restoring the remaining formants as participants showed mastery of the contrast. Using this approach, we have achieved some partial success. First of all, participants generally improved on the training task itself, though with considerable within- and between-participant variability. Many participants improved to the extent that they could reliably identify /r–l/ full-spectrum synthetic speech, even, in some cases, when the tokens were differentiated by a very small difference in F3 onset frequency.

While most participants showed progress in the training task itself, only a subset showed clear improvement from pre- to post-test. Three of six participants trained with feedback and one of six trained without feedback showed a clear pattern of progress in identification of /ra–la/ stimuli differing only in F3. Three of these same participants showed signs of improvement discriminating across the /ra–la/ category boundary. One showed generalization of an acquired NE-like identification function to the /ræ–læ/ category, and two showed improvement in identifying natural /r–l/ minimal pairs produced by several talkers. Given these signs of a training effect in the current experiment, and the near-total absence of a training effect in Experiment 1, it seems natural to infer that our use of an adaptive training paradigm starting with F3 in isolation was a factor in producing the training effect.

The lack of reliable generalization for most participants may be due to limitations in the training stimuli, which lacked variability in non-F3 dimensions. It may be that participants learned to rely on specific conjunctions of cues, i.e. conjunctions of specific F3 onset values with training stimulus values on non-varying dimensions, when identifying the training stimuli. Reliance on these conjunctions, though effective in training, would not result in generalizable reliance on F3 onset frequency and would not result on improved identification on untrained series or natural speech, consistent with what was seen here. Whatever the cause of the poor generalization of training, the fact that there was improvement for some participants from pre- to post-test in this experiment but not Experiment 1 is somewhat encouraging. The finding suggests that initiating training with F3 in isolation can induce some sensitivity to variation in F3 onset frequency within a speech context, even though training using only composite training could not.

General discussion

The difficulty NJ listeners have reliably differentiating English /r/ and /l/ is well documented (Iverson et al., 2005; Lively et al., 1993; Miyawaki et al., 1975; Strange & Dittman, 1984). The primary acoustic cue to category membership is F3 onset frequency. Most NJ speakers do not show NE-like weightings to this cue in perception (Yamada & Tohkura, 1990) or production (Lotto et al., 2004), suggesting that failure to differentiate /r–l/ is based primarily on the use of less reliable cues.

Our aim in these experiments was to test whether restricting /r–l/ variance to the F3 cue would result in improved identification and discrimination performance (Iverson et al., 2003). In particular we hoped to see (1) increases in sensitivity to the F3 cue in the training task itself; (2) improvements from pre- to post-test identifying and discriminating stimuli within the trained vowel context; (3) improved pre- to post-test identification in untrained vowel contexts; and (4) generalization to untrained natural speech /r–l/. These hopes were only partially realized.

The fact that improvement even on the trained series was limited to only a few listeners in Experiment 2 is quite striking, given that other training studies have found much more consistent improvement (e.g., Bradlow et al., 1997; Bradlow et al., 1999; Iverson et al., 2005; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991; McCandliss et al., 2002; Strange & Dittman, 1984). Of the previous studies, only McCandliss et al. (2002) is sufficiently similar to allow direct comparisons. There, adaptive training led to improvement on the trained continuum for all participants after only three sessions of about 500 trials. This contrasts with the present study, where we only found signs of improvement on the trained continuum in 1/3 of the participants after eight 500-trial training sessions. This difference in success rate may be the result of differences in the availability of acoustic cues to /r–l/ category membership other than F3, notably F1 transition duration. Such cues were present in all the earlier studies, including the McCandliss et al. study, but not in the one presented here. These alternative cues appear to be easier for NJ listeners to use to differentiate /r–l/ (Aoyama et al., 2004; Hattori & Iverson, 2009; Gordon et al., 2001; Iverson et al., 2005) and learning to rely on these cues may lead to some degree of generalization, even if this reliance cannot support full NE-like performance.

We did see improvement in 4 of the 12 trained listeners' ability to identify /r–l/ in the trained vowel context, and 2 of these 4 showed clear signs of improvement identifying natural speech. These findings are consistent with examinations of NJ listeners' perception of /r–l/ in natural speech. While most listeners do not rely on the F3 cue to differentiate the sounds, there are some listeners who do rely on the F3 cue, and greater reliance on F3 is associated with more reliable /r–l/ identification (Gordon et al., 2001; Hattori & Iverson, 2009; Ingvalson, McClelland & Holt, 2011; Iverson et al., 2005). Bearing in mind that at least some NJ listeners can rely on the F3 cue in the /r–l/ context, we turn to consider broader theoretical and practical issues in light of our findings.

Speech-related issues

In this section, we consider in what sense speech perception differs from other kinds of auditory processing, why there is such a striking reduction in the ability to acquire spoken language distinctions like the /r–l/ contrast in adulthood, and what it is specifically that makes the /r–l/ contrast so difficult for native Japanese speakers. We approach these issues starting from the last and most specific, building toward the first and most general question.

Difficulty of the /r–l/ contrast for NJ adult speakers

As emphasized throughout, there is evidence that difficulty relying on the F3 onset frequency underlies the difficulty NJ adult speakers have distinguishing English /r/ and /l/. A reasonable conclusion might be that progress in /r–l/ differentiation is achieved through learning to rely on other, less reliable, cues, thus accounting for the incomplete success other training studies have had in producing a robust, English-like ability to distinguish the English /r/ and /l/ phonemes.

While this might be an adequate empirical summary, it is also very puzzling, since the lack of sensitivity to the F3 contrast is not simply a matter of psychoacoustic insensitivity to the onset of F3. Indeed, our Experiment 2 relies upon the fact that NJ listeners can differentiate the F3 outside of the natural speech context. One might be tempted to conclude that Japanese speakers can perceive the F3 contrast in a "non-speech" mode, but cannot use it for the purposes of speech perception. However, a similar F3 contrast can be used by NJ speakers to differentiate /d–g/ (Mann, 1986) and synthetic /r/ and /l/ stimuli can trigger a compensatory adjustment in the perception of the subsequent phoneme (Mann, 1986). Clearly, F3 onset frequency is not simply irrelevant for NJ speech perception.

Reduced ability to acquire spoken language distinctions like the /r–l/ contrast in adulthood

Loss of sensitivity to non-native contrasts and the development of language-specific speech category effects appear to emerge together within the first year of life (Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992; Werker & Tees, 1984). These changes are language-specific, and therefore experience-dependent. These findings suggest that children are initially sensitive to a range of phonemic distinctions and that their sensitivity increases for some distinctions in some contexts but decreases for other distinctions and/or other contexts as a function of the child's early speech experience.

Within this broad context the question remains why the English /r–l/ distinction is of such special difficulty for native Japanese speakers. A number of investigators have researched this issue (MacKain et al., 1982; see Guion et al., 2000 for discussion). One possibility (Flege, 2002, 2003) is that English /r/ and /l/ stimuli activate a strong attractor or perceptual magnet (Kuhl, 1991) associated with the Japanese /r/ phoneme, strongly distorting perception toward /r/ such that both /r/ and /l/ result in the perception of /r/. An alternative (though not necessarily mutually exclusive) possibility arises for the observation that in the most immediately adjacent parts of Japanese phonological space – the space occupied by /r/ – the F3 onset frequency varies freely, making it irrelevant for speech perception (Lotto et al., 2004). One way of coping with such irrelevant variation might be to learn to ignore it, and perhaps such learned ignoring is difficult to reverse (Holt & Lotto, 2006). The prediction arising from this proposal would be that in other cases where there is variation along a dimension that is truly random within a speaker's L1 this should also lead to a highly persistent inability to learn to attend to the dimension of variation if it is used contrastively in a language the speaker attempts to acquire in adulthood.

In what sense if any does speech perception differ from other kinds of auditory processing?

We perceive spoken language to be the product of a human speaker, recognizing particular sounds and words; we hear other kinds of sounds as beeps, chirps, clicks, buzzes, screeches, rumbles, etc. as the products of other sources (animals, tools, musical instruments, etc.). The contrast between these percepts is quite stark phenomenologically, and is clearly notable if one listens to the training stimuli used in Experiment 2. The isolated F3 transitions sound something like a kind of whistle or a chirp, but when the transitions are combined with the other formants at full intensity, native English speakers clearly hear a male voice saying "rah" or "lah". This kind of phenomenology makes it seem as though there may be a special speech mode of perception, quite different from a non-speech mode. While this is a possibility, we suggest there may be a degree of continuity between speech and non-speech. Speech and non-speech sounds are, to be sure, perceived as different things, but the same is also true of the sounds of trumpets and harmonicas. While isolated F3 formants may be more like non-speech than speech, this does not necessarily mean that there is a fundamental mechanistic discontinuity. Reasons to hold this view include the finding that there is categorical perception of non-speech contrasts (Cutting & Rosner, 1976; Mirman, Holt & McClelland, 2004); the finding that animals who are without speech exhibit speech-like processing of some contrasts and can be trained to acquire other contrasts (Kluender, Lotto & Holt, 2005; Kuhl & Miller, 1975; Kuhl & Padden, 1982, 1983; Lotto, Kluender & Holt, 1997a, b), the finding that non-speech adaptors can influence the perception of speech (Holt, 2005; Holt & Wade, 2004; Lotto, Holt & Kluender, 1997; Stephens & Holt, 2003), the finding that there is increasing activation in both size and magnitude at the neural level as the stimulus progresses in complexity from simple non-speech tones to complex non-speech to full speech (Vouloumanos, Kiehl, Werker & Liddle, 2001), and finally, the finding that there are sensitive period effects in the non-speech processing of sounds that parallel many of the findings on sensitive periods in speech processing (Knudsen, 2004; Knudsen & Knudsen, 1990).

Our findings in the present studies are consistent with continuity between speech and non-speech processing mechanisms. In support of this, we found that starting from stimuli that are not perceived as speech, we were able to have an impact on perception of both synthetic and natural speech stimuli, at least in a small subset of cases.

Training and individual differences

It may be worth noting that the total amount of training in our study still occupied only about four hours over a relatively short time period. Even if children learning English as a native language are only exposed to 500 tokens each of /r/ and /l/ per day, by the time they are six months old they will have heard nearly 100,000 tokens of each sound, 50 times as many exposures to tokens of each sound as the participants in our experiments. Thus, the relatively modest signs of progress in our study should not be seen as strong evidence that learning is impossible, or even drastically slowed, in adulthood. Indeed, the fact that any progress occurred with just 4,000 trials of exposure leaves open the possibility that more extended training could result in a greater overall benefit; a change to spaced practice might also produce a greater benefit than what was seen here but in a comparable amount of training trials (Orr & Friedman, 1968).

Pisoni and colleagues (Lively et al., 1993; Lively et al., 1994; Logan et al., 1991) produced a robust, generalizable, and long-lasting training effect (Bradlow et al., 1999) using a wide range of /r–l/ minimal pairs produced by several different speakers, although post-training performance still fell well below that of native English speakers. One way of understanding these findings is to think that participants learn many very specific discriminations during training, each appropriate for a different region of a very complex auditory cue space. Perhaps this set of specific learned discriminations provides broad enough coverage so that there is often a subset that can be drawn upon when novel tokens are encountered. This way of thinking is consistent with the suggestion offered above, that failure of generalization for some of our participants may have resulted from learning the specific combinations of cue values associated with our training stimuli.

An even narrower form of learning may have permitted some participants to progress during training but still to fail to show improvements in identification of stimuli on the /ra–la/ continuum on the post-test. If, during training, listeners are able to retain a memory for the previous stimulus they may be able to compare each present stimulus with the preceding stimulus in order to determine the response. For example, if the previous stimulus had been identified as "r" and the current stimulus is very different from the preceding, that stimulus is likely to be an instance of /l/. While this strategy could be of some utility during training, it would prove less useful during the pre- and post-test, and even less useful when identifying natural speech, where all items differ greatly from one another.

From the outset of the present work, we have emphasized that F3 onset frequency is the most reliable cue to English /r–l/ category membership, and we have hypothesized that robust learning that generalizes might be found if listeners could be induced to rely on the F3 cue. A failure to generalize could occur, however, if our policy of fixing the values of all cues other than F3 in the stimuli led some participants to learn conjunctions of cues that differentiated the training stimuli but would not be available in our untrained synthetic speech stimuli or in natural speech. However, we note again that at least a subset of participants was able to learn in a way that generalized, at least to some degree. This suggests that, for some participants at least, our training procedure was successful in inducing a reliance on the F3 cue that extends beyond the narrow confines of the training context.

In both the present effort and in previous work examining NJ perception of /r–l/, there is extensive evidence of individual differences. Of particular interest here, 14 individuals across the two experiments were excluded from training because they were already sensitive to F3 onset before training, and 4 of 12 trained individuals in Experiment 2 showed a benefit of training focusing on F3. Though many NJ listeners do not rely on F3 onset frequency to differentiate /r–l/, it is clear that some NJ individuals do show reliance on this cue (Gordon et al., 2001; Hattori & Iverson, 2009; Ingvalson et al., 2011). Thus, it is clear that the inability of NJ listeners' to distinguish the English /r/ and /l/ on the basis of the F3 cue is neither total nor uniform. Furthermore, recent efforts to tie neuroimaging data to second-language training have revealed that individual differences in neuroanatomy and neurophysiology can predict how successful a language learner might be (Golestani, Molko, Sehaene, LeBihan & Pallier, 2007; Wong, Perrachione & Parrish, 2007; Wong, Warrier, Penhune, Roy, Sadehh, Parrish & Zatorre, 2008). For example, Raizada and colleagues found that the statistical separability between patterns of activation to /ra/ and /la/ predicted individuals' abilities to behaviorally distinguish the sounds (Raizada, Tsao, Liu & Kuhl, 2009). Similarly, Zhang, Kuhl, Imada, Iverson, Pruitt, Stevens, Kawakatsu, Tohkura, and Nemoto (2009) found that all listeners showed changes in neural and behavioral responses following training and that the degree of change between the neural and behavioral responses were correlated. Further work on the basis of individual difference in NJ speakers' ability to rely on F3 will be needed to uncover just what it is that allows some NJ speakers to use this cue effectively.

Conclusion

We are a long way from a clear understanding of how it might be possible to produce a training regime that could lead NJ listeners to identify /r–l/ at NE-like rates. To date, studies using high variability natural speech training have produced the most generalizable and long-lasting effects. Success with such methods is still incomplete, however, and may reflect a failure to learn to rely on the F3 transition cue. Our efforts to teach such reliance in the present study were at best a limited success, but to us there were enough promising signs that our approach might be of use as a starting place for future investigations. It may be that the development of a fully successful strategy for teaching the /r–l/ distinction to NJ speakers will never be found, but it is clear that no such strategy will be found without further experimental investigations. We encourage others to join in these efforts.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7,* 131–153.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32,* 233–250.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics, 61,* 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299–2310.

Cutting, J. E., & Rosner, B. S. (1976). Discrimination functions predicted from categories of speech and music. *Perception & Psychophysics, 20,* 87–88.

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English. *Journal of the Acoustical Society of America, 92 (2),* 736–757.

Flege, J. E. (2002). Interactions between the native and second-language phonetic systems. In P. Burmeister, T. Piske & A. Rhode (eds.), *An integrated view of language development: Papers in honor of Henning Wode*, pp. 217–244. Trier: Wissenschaftlicher Verlag Trier.

Flege, J. E. (2003). Assessing constraints on second language segmental production and perception. In N. O. Schiller & A. Meyer (eds.), *Phonetics and phonology in language comprehension and production, differences and similarities*, pp. 319–355. Berlin: Mouton de Gruyter.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults' perception of /r/ and /l/. *Journal of the Acoustical Society of America, 99,* 1161–1173.

Golestani, N., Molko, N., Sehaene, S., LeBihan, D., & Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cerebral Cortex, 17,* 575–582.

Gordon, P. C., Keyes, L., & Yung, Y.-F. (2001). Ability in perceiving nonnative contrasts: Performance on natural and synthetic speech stimuli. *Perception & Psychophysics, 63,* 746–758.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107,* 2711–2724.

Hattori, K., & Iverson, P. (2009). English /r/–/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *Journal of the Acoustical Society of America, 125 (1),* 469–479.

Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science, 16,* 305–312.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059–3071.

Holt, L. L., & Wade, T. (2004). Non-linguistic sentence-length precursors affect speech perception: Implications for speaker and rate normalization. In Slifka et al., pp. C49–C54.

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics, 39,* 571–584.

Ingvalson, E. M., & Wenger, M. J. (2005). A strong test of the dual mode hypothesis. *Perception & Psychophysics, 67,* 14–35.

Iverson, P., Ekanayake, D., Hamann, S., Sennema, A., & 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. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1305–1316.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r–l/ to Japanese adults. *Journal of the Acoustical Society of America, 118,* 3267–3278.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* B47–B57.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/–/θ/ contrast by francophones. *Perception & Psychophysics, 40,* 205–215.

Jenkins, J. J., Strange, W., & Polka, L. (1995). Not everyone can tell a "rock" from a "lock": Assessing individual differences in speech perception. In D. Lubinski & R. V. Dawis (eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings*, pp. 297–325. Palo Alto, CA: Davies-Black.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21,* 60–99.

Klatt, D. H. (1980). Software for a cascade-parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971–995.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820–857.

Kluender, K. R., Lotto, A. J., & Holt, L. L. (2005). Contributions of nonhuman animal models to understanding human speech perception. In S. Greenberg & W. Ainsworth (eds.), *Listening to speech: An auditory perspective*, pp. 203–220. New York: Oxford University Press.

Knudsen, E. I. (2004). Sensitive periods in the development of brain and behavior. *Journal of Cognitive Neuroscience, 16,* 1412–1425.

Knudsen, E. I., & Knudsen, P. F. (1990). Sensitive and critical periods for visual localization of sound calibration by barn owls. *Journal of Neuroscience, 10,* 222–232.

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perceptual Psychophysics, 50,* 93–107.

Kuhl, P. K. (1993). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In B. deBoysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage & J. Morton (eds.), *Developmental neurocognition: Speech and face processing in the first year of life*, pp. 259–274. Dordrecht: Kluwer.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced–voiceless distinction in alveolar plosive consonants. *Science, 190,* 69–72.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics, 32,* 542–550.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America, 73,* 1003–1010.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: John Wiley & Sons.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358–368.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94,* 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify

English /r/ and /l/: III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96,* 2076–2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874–885.

Lotto, A. J., Holt, L. L., & Kluender, K. R. (1997). Effect of voice quality on perceived height of English vowels. *Phonetica, 54,* 76–93.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997a). Animal models of speech perception phenomena. In K. Singer, R. Eggert, & G. Anderson (eds.), *Chicago Linguistic Society* (vol. 33), pp. 357–367. Chicago: Chicago Linguistic Society.

Lotto, A. J., Kluender, K. R., & Holt, L. L., (1997b). Perceptual compensation for coarticulation by Japanese quail (*Coturnix cotrunix japonica*). *Journal of the Acoustical Society of America, 102,* 1134–1140.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: Case of Japanese acquisition of /r/ and /l/. In Slifka et al. (eds.), pp. C181–C186.

MacKain, K. S., Best, C. T., & Strange, W. (1982). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2,* 369–390.

Maddox, W. T., Diehl, R. L., & Molis, M. R. (2001). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. In R. Smits, J. Kingston, T. M. Nearey & R. Zondervan (eds.), *Proceedings of the Workshop on Speech Recognition as Pattern Recognition,* pp. 85–90. Nijmegen: MPI for Psycholinguistics.

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r". *Cognition, 24,* 169–196.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Predictions of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective and Behavioral Neuroscience, 2,* 89–108.

Mirman, D., Holt, L. L., & McClelland, J. M. (2004). Categorization and discrimination of non-speech sounds: Differences between steady-state and rapidly-changing acoustic cues. *Journal of the Acoustical Society of America, 116,* 1198–1207.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. L., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Attention, Perception, & Psychophysics, 18,* 331–340.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word, 13,* 24–43.

Orr, D. B., & Friedman, H. L. (1968). Effect of massed practice on the comprehension of time-compressed speech. *Journal of Educational Psychology, 59,* 6–11.

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *Journal of the Acoustical Society of America, 78 (4),* 1187–1197.

Raizada, R. D. S., Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2009). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: Prediction of individual differences. *Cerebral Cortex, 20 (1),* 1–12.

Romaine, S. (2003). Variation. In C. J. Doughty & M. H. Long (eds.), *The handbook of second language acquisition*, pp. 409–435. Oxford: Blackwell.

Slifka, J., Manuel, S., & Matthies, M. (eds.) (2004). *Proceedings of From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*. Cambridge, MA: MIT Press.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech. *Journal of the Acoustical Society of America, 114,* 3036–3039.

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r–l/ by Japanese adults learning English. *Perception & Psychophysics, 36,* 131–145.

Takagi, N. (2002). The limits of training Japanese listeners to identify English /r/ and /l/: Eight case studies. *Journal of the Acoustical Society of America, 111,* 2887–2894.

Takagi, N., & Mann, V. (1995). The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of English. *Applied Psycholinguistics, 16,* 379–405.

Underbakke, M., Polka, L., Gottfried, T. L., & Strange, W. (1988). Trading relations in the perception of /r/–/l/ by Japanese learners of English. *Journal of the Acoustical Society of America, 84 (1),* 90–100.

Vouloumanos, A., Kiehl, K. A., Werker, J. F., & Liddle, P. F. (2001). Detection of sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and nonspeech. *Journal of Cognitive Neuroscience, 13 (7),* 994–1005.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7,* 49–63.

Wong, P. C. M., Perrachione, T. K., & Parrish, T. B. (2007). Characteristics of successful and less successful speech and word learning in adults. *Human Brain Mapping, 28,* 995–1006.

Wong, P. C. M., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., & Zatorre, R. J. (2008). Volume of left Heschl's gyrus and linguistic pitch learning. *Cerebral Cortex, 18,* 828–836.

Yamada, R. A., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. *Proceedings of the 1990 International Conference on Spoken Language Processing*, pp. 757–760. Kobe, Japan.

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., Kawakatsu, M., Tohkura, Y., & Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage, 46,* 226–240.