Cognition 115 (2010) 435–443

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Unattended exposure to components of speech sounds yields same benefits as explicit auditory training

Aaron R. Seitz [a,b,*,1], Athanassios Protopapas [c,1], Yoshiaki Tsushima [a,e,1,2], Eleni L. Vlahou [c,d], Simone Gori [f,2], Stephen Grossberg [a,g], Takeo Watanabe [a,g]

[a] Center of Excellence for Learning in Education, Science and Technology, 677 Beacon st, Boston, MA 02215, USA
[b] University of California, Riverside, USA
[c] Institute for Language & Speech Processing, "Athena" Research Center, Greece
[d] Department of Psychology, University of Crete, Rethimno, Greece
[e] Harvard University, William James Hall, 33 Kirkland St., Cambridge, MA 02138, USA
[f] Department of General Psychology, University of Padua, Italy
[g] Department of Cognitive and Neural Systems, Boston University, USA

ARTICLE INFO

Article history:
Received 9 June 2009
Revised 13 February 2010
Accepted 1 March 2010

Keywords:
Perceptual learning
Audition
Formants
Speech
Implicit learning

ABSTRACT

Learning a second language as an adult is particularly effortful when new phonetic representations must be formed. Therefore the processes that allow learning of speech sounds are of great theoretical and practical interest. Here we examined whether perception of single formant transitions, that is, sound components critical in speech perception, can be enhanced through an implicit task-irrelevant learning procedure that has been shown to produce visual perceptual learning. The single-formant sounds were paired at subthreshold levels with the attended targets in an auditory identification task. Results showed that task-irrelevant learning occurred for the unattended stimuli. Surprisingly, the magnitude of this learning effect was similar to that following explicit training on auditory formant transition detection using discriminable stimuli in an adaptive procedure, whereas explicit training on the subthreshold stimuli produced no learning. These results suggest that in adults learning of speech parts can occur at least partially through implicit mechanisms.

1. Introduction

Languages differ in their phonetic repertoire, that is, in the set of speech sounds that are used to form words and thus to convey distinctions in meaning. Infants learn the speech sounds of their linguistic environment in their first year of life by attending to sound differences that are related to meaning differences and ignoring inconsequential sound differences (Jusczyk, 1997). This results in more efficient processing of speech sounds used in their language and less efficient processing of other sounds (Kuhl et al., 2008). Language acquisition, in general, and phonetic learning, in particular, appear to rely heavily on implicit learning mechanisms that extract statistical regularities organized at many different levels (Perruchet & Pacton, 2006; Saffran, Werker, & Werner, 2006). For example, humans' sensitivity to the distributional frequencies of the acoustic input affects word segmentation and phonetic categorization (Maye, Werker, & Gerken, 2002; Saffran, Newport, & Aslin, 1996). These powerful statistical mechanisms are modulated by attentional and motivational factors (Kuhl, Tsao, & Liu, 2003) as well as contingent positive reinforcements (Goldstein, King, & West, 2003;

* Corresponding author. Address: Department of Psychology, UC Riverside, 900 University Ave., Riverside, CA 92521, USA. Tel.: +1 951 827 6422.
E-mail address: aseitz@ucr.edu (A.R. Seitz).
[1] ARS, AP, and YT are equal contribution authors.
[2] YT and SGo conducted this research at Boston University.

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.03.004

Gros-Louis, West, Goldstein, & King, 2006). However very little is known regarding the mechanisms that guide phonetic learning in adults.

Despite initial nondiscriminability, adults can learn to distinguish new phonetic contrasts (for review see Bradlow (2008) and Pisoni, Lively, and Logan (1994)). Substantial and long-lasting gains are seen (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994), generalizing to some extent to production (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997), though this learning is limited to achieving performance levels well below native levels. A well-studied example concerns learning of the English /r/–/l/ distinction by Japanese adults. American English /r/ differs mostly from /l/ in the frequency of the third spectral peak (third formant or F3; see Fig. 1), which is very low for /r/ but as high as possible for /l/ (Stevens, 1998). Although this is not the only acoustic difference between /r/ and /l/, variation in F3 onset and transition is sufficient for native speakers of American English to discriminate between /r/ and /l/ (O'Connor, Gerstman, Liberman, Dalattre, & Cooper, 1957; Yamada & Tohkura, 1990). Also, Japanese listeners who are unable to discriminate English /r/ from /l/ exhibit difficulty in differentially processing F3 in the acoustic context of a syllable (Yamada & Tohkura, 1990).

Phonetic training regimes for adults differ dramatically in their methods and in their underlying assumptions regarding the mechanisms involved in learning. Phonetic learning has been found through explicit phonetic training with focused attention on the stimulus differences, explicit category labels, and performance feedback (Bradlow, 2008; Loebach & Pisoni, 2008; Pisoni et al., 1994; Vallabha & McClelland, 2007). It is also seen under natural settings after prolonged experience in a non-native phonetic, linguistic and social environment (Flege, 2003), where learning of the critical differences that distinguish phonetic contrasts emerges largely unintentionally. Phonetic training studies generally employ explicit training procedures, with participants focusing their attention on distinguishing the phonetic contrasts and receiving response feedback (Loebach & Pisoni, 2008; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). Some degree of learning without external feedback is possible when stimuli are made discriminable through exaggeration (McCandliss et al., 2002), a finding consistent with Hebbian learning mechanisms (Grossberg, 1978, 1987; Gutnisky, Hansen, Iliescu, & Dragoi, 2009; Vallabha & McClelland, 2007) reinforcing the distinct percepts produced by exaggerated stimuli (Vallabha & McClelland, 2007). However, a reliably larger gain and more rapid improvement was found in training with feedback (compared to training without feedback), indicating that the simple Hebbian-learning account is "at best, incomplete" (McCandliss et al., 2002, p. 104).

Here we examine how novel approaches to perceptual learning may shed light on the mechanisms involved in adult phonetic learning. We consider a recent model of task-irrelevant perceptual learning (TIPL) (Seitz & Watanabe, 2005), which views perceptual learning as the result of systematic coincidences between: (a) stimulus-driven representations upon exposure to environmental stimuli and (b) diffuse signals elicited upon successful task performance. In this model, stimulus features are represented and available for reinforcement learning whether attended or not. This representation is pre-perceptual in that it may occur below limens of detectability or discriminability. The "success signals" that modulate learning may be elicited by external rewards (Seitz, Kim, & Watanabe, 2009) or by internally generated performance evaluation in lieu of feedback (Seitz & Watanabe, 2009). A key prediction of this model is that in the course of performing a task, the individual learns unattended stimulus features, in addition to attended stimuli, that coincide with successful performance, because the modulating signal is not tied to the specific stimulus features causing its elicitation. This model is consistent with neural models of learning, attention, and motivation during reinforcement learning (Dranias, Grossberg, & Bullock, 2008; Grossberg & Merrill, 1996), while it stands in contrast to frameworks in which

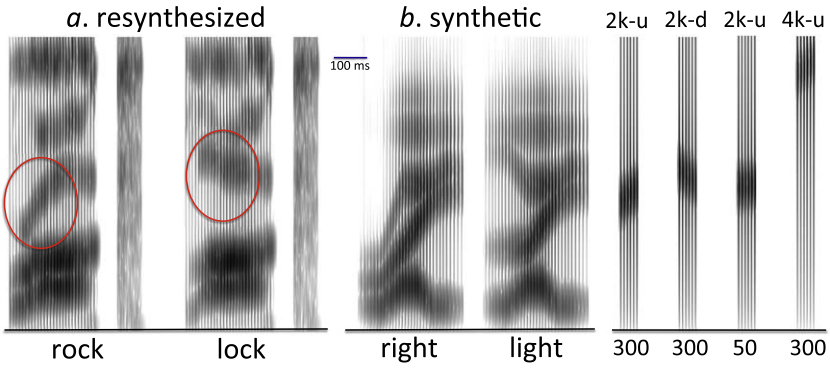

Fig. 1. English /r/–/l/ distinction and examples of our formant transition stimuli. *Left:* spectrograms of LPC-resynthesized "lock" and "rock" stimuli based on natural recordings, from McCandliss et al. (2002). These sounds differ mainly in the 3rd spectral prominence (formant; circled in red). *Middle:* spectrograms of synthetic "right" and "light" stimuli (excluding the final /t/burst) based on the specifications of Yamada and Tohkura (1990). Note the initial steady-state and transition difference in the 3rd formant. *Right:* spectrograms of synthesized single formant transitions used in our study, including examples from different conditions. 2k/4k refers to 2600 Hz vs. 4600 Hz endpoint center frequency; u/d refers to upward vs. downward sweep direction; 50/300 refers to the extent of the transition sweep in Hz within the 70-ms stimulus duration. All spectrograms are plotted to the same scale, extending from 0 to 5000 Hz in the vertical direction; the common horizontal scale is indicated by a 100-ms segment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

learning is gated by task-directed attentional factors (Ahissar & Hochstein, 1993).

The present work extends the TIPL procedure into the auditory domain, addressing, in particular, the sound property that is most important for distinguishing /r/ from /l/. We used subthreshold single formant transitions as unattended, task-irrelevant, stimuli that were presented in a temporally correlated manner within sequences of task-relevant animal sounds (see Fig. 2 for task schematic). We found that after 10 days of training on the serial auditory presentation (SAP) animal sound identification task, subjects improved at discriminating formant transitions that had been temporally paired with targets of the SAP task. Notably, the magnitude of the threshold improvements found from the TIPL procedure was comparable to that achieved through explicit training with feedback for the same auditory distinction.

2. Method

2.1. Participants

Thirty-two adults (18–35 years old), with normal hearing and normal or corrected-to-normal vision, participated in the study. In the TIPL training, 16 subjects participated, four in each of four conditions. Of these, eight were native English speakers, six native Japanese speakers (one had an English speaking parent), and two native Chinese speakers. In the adaptive training, eight subjects participated, four in each of two conditions, including five native English speakers, one native Japanese speaker, one native Korean speaker, and one native Italian speaker. In the explicit training, eight more subjects participated, four in each of two conditions, including four native English speakers and four native Chinese speakers. One participant in the explicit

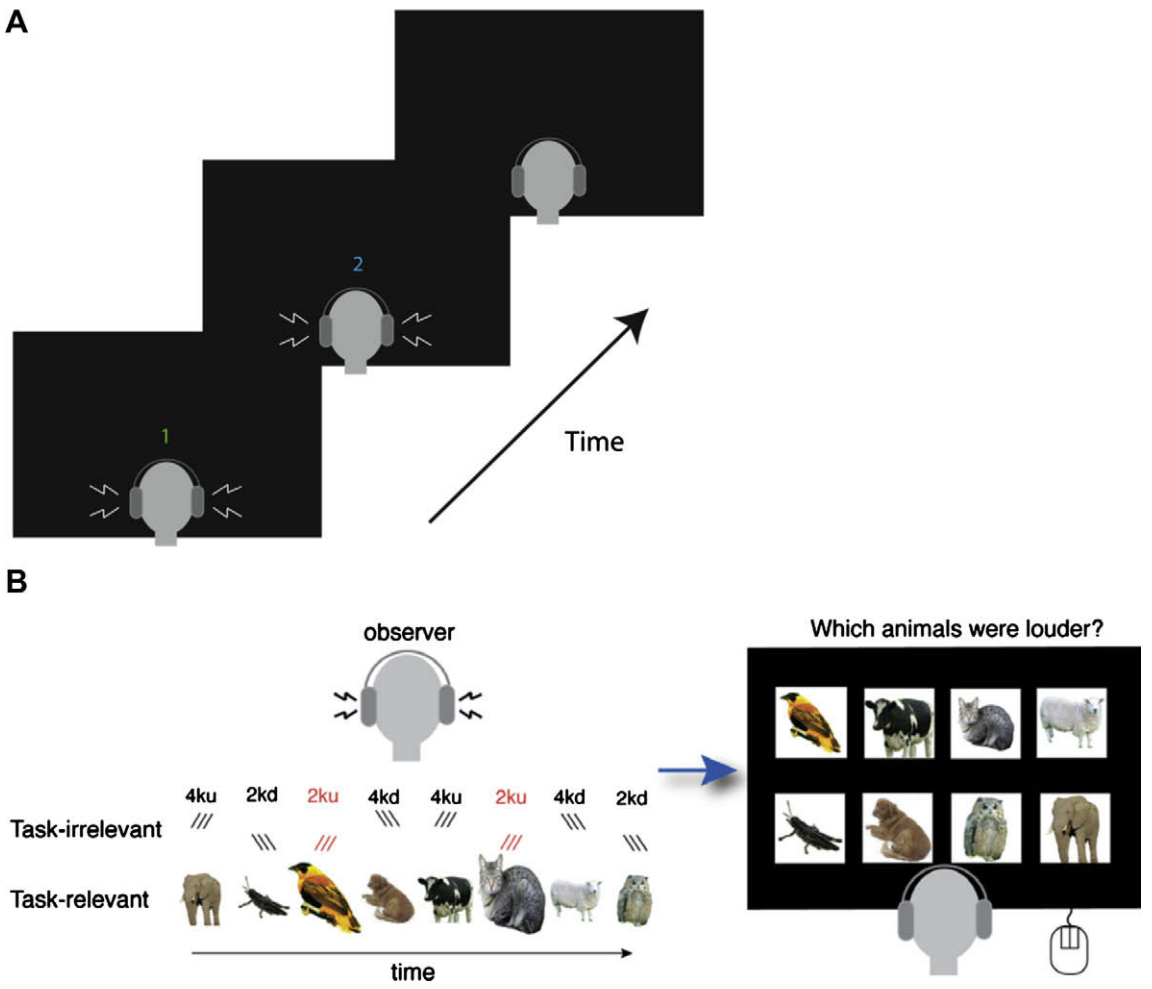

Fig. 2. Task schematics. (A) Schematic of two-interval formant transition detection task; observer heard two sounds and had to report whether the first or second sound contained the formant sweep. (B) Schematic of SAP task; observer heard eight sounds and had to click on the two pictures that corresponded to the two louder sounds (indicated in cartoon by larger animals; bird and cat) in the sequence. Task-irrelevant formant transition stimuli were presented three times during each animal sound, as indicated by triple lines (/for upward sweeps and/for downward sweeps; with higher elevation indicating higher frequencies), with formant transitions paired with task-targets shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

training group for 2ku was dropped from the study due to his extremely poor performance on the pre-test (initial average threshold of 800 Hz in the 2ku condition) and thus only seven participants were included in the explicit training data set reported below. We observed no differences in learning based upon language background, as non-native English speakers in each condition showed qualitatively similar results as the native English speakers in the same conditions. Participants were naïve as to the purpose of the experiments. Informed consent was obtained according to the requirements of the Boston University and University of California at Riverside, Institutional Review Boards.

2.2. Stimuli

Formant transitions were 70 ms long and were created by passing a constant-amplitude train of seven impulses spaced 10 ms apart through a simple resonator with center frequency continuously varied (linearly interpolated for every sample at a rate of 22,050 Hz) from an initial value (specified below) to a final value of either 2600 or 4600 Hz and a constant bandwidth of 260 or 460 Hz, respectively, resulting in sounds with a single spectral peak transition at a constant fundamental frequency of 100 Hz and an approximate intensity of 63.7 dBA SPL. Resonating frequencies at stimulus onset were determined following transition detection threshold estimation (see below). For the training, onset frequencies were set at subthreshold values of 2475 (condition "2k-up"), 2725 (2k-down), 4400 (4k-up), and 4800 (4k-down) Hz for the first eight participants. For the second eight participants values were chosen at ~80% of the threshold level of the first group resulting in values of 2500 (2k-up), 2750 (2k-down), 4300 (4k-up), and 5100 (4k-down). Thus the total extent of the formant sweep over the 70 ms was 125 (100), 125 (150), 200 (300), and 200 (500) Hz, for the 2k-up, 2k-down, 4k-up, and 4k-down conditions, respectively (values for the second group in parenthesis). Reference constant-profile single-formant stimuli for each condition were constructed with a constant resonating frequency equal to the fixed endpoint value of the corresponding set (2600 Hz for 2k and 4600 Hz for 4k).

Stimuli for the SAP task were eight identifiable animal sounds (dog, cat, cow, bird, elephant, sheep, cricket, and owl) downloaded from http://www.seaworld.org/animal-info/sound-library/index.htm (Supplemental Fig. 4 displays spectrograms of these sounds). Sounds were equated in duration at 500 ms, using 10-ms square-sine on and off ramps as needed, and in intensity, to the extent possible, at approximately 62.4 dBA SPL.

2.3. Procedure

The experiment consisted of four phases: practice, pre-test, training sessions, and post-test. In practice sessions, participants were familiarized with the formant transition stimuli and with the psychoacoustic threshold estimation procedure. There were two endpoint-frequency conditions (termed 2k and 4k; see "Stimuli") crossed with two sweep-direction conditions (up and down), resulting in four detection conditions, each run once during practice. The pre-test was conducted on the day following practice. In this session, formant transition detection thresholds were estimated four times (in nonconsecutive runs) for each condition. In the 10 training sessions, taking place over 10 consecutive days, participants carried out the SAP animal sound identification task. Finally, the post-test was identical to the pre-test.

2.3.1. Formant transition detection task (practice and test sessions)

Formant transition detection thresholds were determined psychoacoustically in a two-interval forced-choice task presenting one formant sweep and one stimulus with a constant spectral profile (fixed resonator) at each trial, in random order. Participants pressed "1" or "2" on the keyboard to indicate the interval of the sweeping stimulus. Stimuli differed only in resonator center frequency, determined adaptively in a modified variable-step staircase procedure based on accelerated stochastic approximation (Treutwein, 1995), with $c = 400$ (starting step size 250 Hz and initial sweep extent of 600 Hz) and target correct response probability .75. The procedure was terminated at 15 reversals unless a maximum of 60 trials was reached first. Thresholds were calculated by linear averaging of the extent of formant sweep for the last 12 response reversals (or as many as available, if fewer). Thresholds were determined in separate runs for 2k-up, 2k-down, 4k-up, and 4k-down conditions. Accuracy feedback was given in the practice session only.

2.3.2. TIPL training sessions

The TIPL training involved SAP of eight animal sounds with 50 ms interstimulus interval (ISI). After each trial, participants reported the two louder animal sounds in the sequence by clicking (in the correct order) on two of eight animal pictures displayed on the screen (Fig. 2). The order of eight animal sounds and the two animals chosen as targets was randomized across trials. A two-component adaptive procedure was applied to the amplitude of the two target sounds during this task. First, an adaptive staircase affected the amplitude increment of the target items (after each correct trial dB increment over base was multiplied by .95 and after incorrect trials divided by .9025). Due to extreme differences in spectral profiles among the animal sounds, intensity levels were adaptively varied independently based on identification performance for each animal separately. Thus, a second staircase applied a multiplier for each individual animal sound (dB increment multiplied by .995 or divided by .9925). This procedure achieved consistent performance across participants and animal sounds (Supplemental Fig. 2). No accuracy feedback was given.

Throughout training, formant transition sounds were presented simultaneously (linearly added) with the animal sounds. For each participant, a specific subthreshold formant transition (2ku, 2kd, 4ku, or 4kd) was added to the targets of every trial (paired sound), and the other three formant sweeps were added to the distractors (nonpaired sounds), such that the four different formant transitions were presented an equal number of times in each trial.

The choice of paired sound was counterbalanced across participants. Given that the animal sounds were presented for 500 ms each and the formant transitions were only 70 ms in duration, each formant transition was presented 3 times during a single animal sound presentation, with a 72.5 ms ISI between formant presentations. During the 10 days of training, each participant heard the paired formant sounds three times with two animals in each trial, totaling 18,000 individual presentations of the subthreshold single-formant sweep over 3000 trials (300 trials per session).

The base level of the animal sounds was approximately 62.4 dBA and the paired formant stimuli were presented at 55.1 dBA with combined intensity of the formant + animal sound of about 63.4 dBA for nontargets. For targets, the animal component of the sound was elevated by the aforementioned adaptive procedure as follows: $I_t = I_f + (1.0 + -t_o \times t_a) \times I_a$ (I_f: formant sweep level; I_a: animal sound base level; t_o: overall threshold; t_a: animal sound-specific threshold).

2.3.3. Explicit training sessions

Explicit training was designed to test the possibility that the subthreshold stimuli are learnable when attended and externally rewarded. The single-formant stimulus stream was based on that used in TIPL training, thus consisting of two target triplets and six nontarget triplets. The two target triplets (which were paired with the louder animal sounds in TIPL) were the same formant transitions in each trial as they were in TIPL. The other six stimuli were sets of three identical single-formant sounds with constant spectral profiles, their formant frequency being fixed at the endpoint frequency of the corresponding stimuli used in TIPL, thus being identical to the constant-profile reference stimuli used in testing. Only two frequency conditions were used in explicit training (2kd and 2ku, counterbalanced across subjects).

To minimize confusion of the participants, we removed the animal sounds from the auditory stream and replaced them with a sequence of animal pictures. The participants' task was to report the two animals that were paired with the formant transitions, using the same response screen as in TIPL. Due to the unchanged structure of presentation of the target stimuli, each subthreshold single-formant sweep was presented 18,000 times, as many as in TIPL, in the same number of trials.

2.3.4. Adaptive training sessions

Procedures for standard adaptive training were similar to the other training conditions in terms of scheduling and testing, and similar to the test sessions in terms of the stimulus–response interaction. Participants conducted the same practice and pre- and post-training test sessions, however only the 2ku and 2kd sounds were used. Each participant conducted five training sessions on separate days in which either the 2ku or 2kd formant was trained (four participants in each group). The training was identical to the test sessions with the exception that response accuracy feedback was given after each trial. Participants conducted eight repetitions of the threshold estimation procedure in each training session, for a maximum of 2400 individual presentations of single-formant sweeps and potential corresponding rewards.

3. Results

The results from 16 subjects who conducted the SAP task show that the training task remained difficult and that subjects underwent task-related learning in identifying the loudest animal sounds (thresholds for each day shown in Supplemental Fig. 1). Threshold decrease across sessions was significant by two-way repeated measures ANOVA ($F_{(9,\ 1248)} = 39.18$, $p < .001$). Loudness thresholds for different animal sounds were significantly different ($F_{(7,\ 1248)} = 71.84$, $p < .001$), however identification accuracy was highly similar across animal sounds ($F_{(7,\ 1248)} = 1.18$, $p = .312$) presumably due to the target-specific adaptive procedure (see Supplemental Fig. 2 for performance on each target-type).

Mean formant transition detection thresholds before and after training are shown in Fig. 3. Threshold values were higher than formant sweep extents used during training, confirming that formant transitions presented during training were indeed subthreshold. There were no significant differences in pre-training thresholds between paired and unpaired conditions (paired samples t-test, averaging distance from mean pre-training threshold for each condition, $t_{(15)} = 1.25$, $p = .231$; the same results were obtained comparing ratios to condition means, $t_{(15)} = 1.26$, $p = .226$).

Our main interest was to see whether learning would occur specifically for formant transitions paired with the SAP targets, that is, whether significant differences would emerge between post-training and pre-training thresholds in the paired condition (Fig. 4). A significant decrease was found in detection thresholds for formant transitions paired with targets ($t_{(15)} = -2.3$, $p = .037$, paired samples t-test), but not for formant transitions paired with nontargets ($t_{(47)} = .17$, $p = .87$). There was also a significant difference in learning between formant transitions paired with targets and formant transitions paired with nontargets ($t_{(15)} = -2.75$, $p = .015$, paired samples t-test of paired sound threshold change vs. average of unpaired sounds). These results confirm that task-irrelevant learning occurred for acoustic components that distinguish speech sounds even when the dimension undergoing learning remained unattended and subthreshold.

The results of seven subjects in the explicit training group are plotted in the far right (Trained RSAP 2k) of Fig. 4. We found no evidence that these subjects were able to improve their formant transition detection sensitivity through this procedure, even though they experienced the exact same number of stimulus presentations per day, using the same target stimuli, over the same number of days as the implicit training group.

In the adaptive training group, performance improvements asymptoted within 5 days (Supplemental Fig. 3 shows day-by-day performance). The degree of learning was not significantly different from that found after implicit, task-irrelevant training ($t_{(14)} = .14$, $p = .89$; unpaired t-test of trained 2k threshold changes between groups). Effect sizes (Cohen's d) were 1.1 and 0.9 in the adaptively

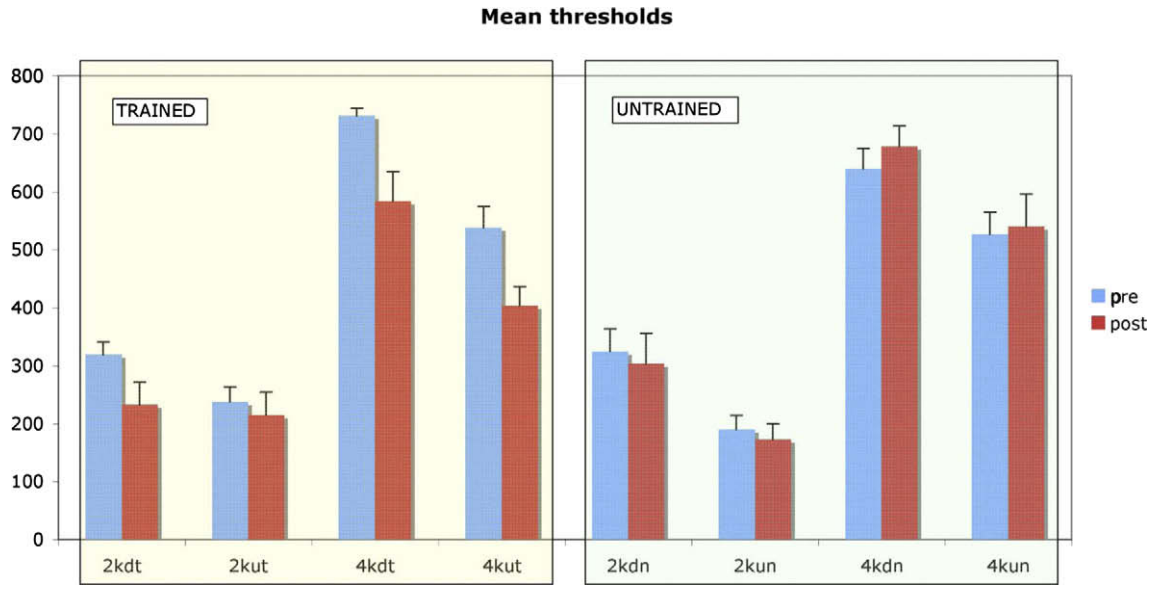

Fig. 3. Formant sweep extent detection performance (mean thresholds, in Hz) for each condition, before and after training. "Trained" (*t*) refers to sweeps paired with target animal sounds during training and "untrained" (*n*) refers to sweeps paired with nontargets (averaged across the corresponding subjects in each case). Error bars show standard error.

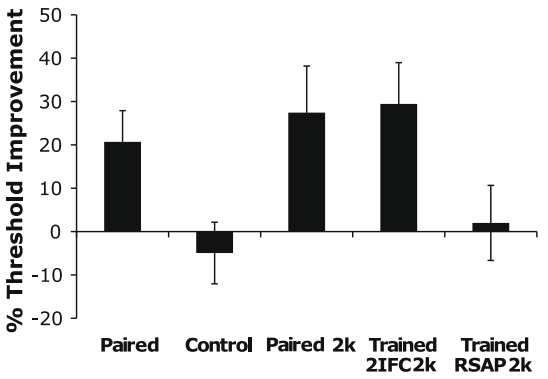

Fig. 4. Change in formant sweep extent detection thresholds between test sessions, before vs. after training, as a proportion (percentage) of pre-training threshold, in Hz. Paired, average change across the 2kd, 2ku, and 4kd, 4ku conditions for formant transitions paired with SAP targets. Control, average threshold change for formant transitions not paired with targets in the SAP task. Paired 2k, average change of 2kd and 2ku thresholds from task-irrelevant training. Trained 2k, average change of 2kd and 2ku thresholds from adaptive training. RSAP 2k, average change of 2kd and 2ku thresholds from explicit training. Error bars show standard error.

and implicitly trained group, respectively (see "trained 2IFC 2k" and "paired 2k" in Fig. 4). As with implicit training, performance benefits were specific to the trained formant transition in the adaptive training group ($t_{(7)} = -2.8$, $p = .026$, paired samples *t*-test of trained sound threshold change vs. untrained threshold change).

4. Discussion

Our results show that detection thresholds of auditory formant transitions can be lowered implicitly by pairing these sounds with the targets of an unrelated task. Neither attention nor awareness of the critical stimulus property (i.e., change in spectral peak) is necessary to achieve an increase in sensitivity to sweep extents of formant transitions.

While the sweep extents of the formant transitions used during TIPL were below participants' thresholds, the presence of these formant transition stimuli was perceptible. During debriefing at the end of the experiment, some participants noted that they heard "clicking sounds". However, no participant could indicate any relationship between the formant transitions and the animal sound targets. Also while the loudness of the targets was increased relative to the distractors, the formant transitions were presented at constant amplitude, therefore the target-paired formants were presumably more difficult to identify than the others. Furthermore, we used a fully balanced design, in which formant sounds were counterbalanced across participants, performance was controlled with an adaptive staircase, and each formant transition condition was presented simultaneously with each animal sound (both as target and distractor) an equal number of times. These factors give us confidence that learning obtained through our procedure cannot be explained by selective attention to particular formant transitions during the training procedure.

There were two differences in the auditory stimuli between the explicit and implicit group, namely the absence of animal sounds and the constant profile of the nontarget single-formant stimuli in the explicit group. Both of these differences were expected to provide a learning advantage to the explicit group, because there was less interference from other sounds and no confusability of the direction of spectral change, respectively. Yet, there was no evidence for learning in the explicit group, suggesting that explicit

attention may be insufficient for (or even detrimental to; see Choi, Seitz, & Watanabe, 2009) the formation of higher sensitivity representations of the subthreshold stimuli. An additional potential difference is that, in the implicit group, loud animal sounds may have acted as internal feedback generators, reinforcing almost every presentation of the target single-formant sweeps, whereas in the explicit group, reinforcing external feedback occurred only in the rare instances in which participants guessed correctly which were the target stimuli. Therefore both the role and the source of feedback warrant further scrutiny in future work.

While the magnitude of learning was similar between the implicit and adaptive groups, there were substantial differences between the stimuli and the procedures, favoring learning in the adaptive group. Specifically, in adaptive training stimuli were consistently presented at or above discrimination threshold during training, and were individually rewarded at a high rate of performance, whereas in implicit training stimuli were consistently presented at subthreshold levels and there was no specific external reward associated with their presentation. Moreover, the adaptive procedure tracked the participants' performance by adjusting the sweep extent and thus maintaining difficulty in the trained task at a near-optimal level for perceptual improvement (Amitay, Irwin, Hawkey, Cowan, & Moore, 2006). On the other hand, implicit training lasted for 10 days, with more stimulus presentations per day than adaptive training, which only lasted 5 days. However, there was evidence of asymptotic performance in adaptive training, suggesting that further training might have had little or no additional learning effect.

Because of the notorious difficulty Japanese listeners face with the English /r/–/l/ distinction, it has been considered as a testing ground for different non-native phonetic training approaches. An important lesson from such training paradigms is that variability in training is critical for achieving transfer of stimulus-specific learning to new speakers, new sound tokens, and new phonetic environments (e.g. Bradlow, 2008; Lively, Logan, & Pisoni, 1993; McCandliss et al., 2002). Feedback is generally considered necessary for learning. However, when the critical acoustic differences are artificially exaggerated, making them more easily identifiable, learning is possible without performance feedback, as Hebbian learning mechanisms presumably reinforce the distinct percepts that are produced by the exaggerated stimuli, resulting in separate categories (McCandliss et al., 2002; Vallabha & McClelland, 2007). The Hebbian account cannot provide a complete account of phonetic category learning, because feedback markedly improves learning (e.g. Vallabha & McClelland, 2007). Tricomi, Delgado, McCandliss, McClelland, and Fiez (2006) suggested that the strong activation of the caudate nucleus observed in phonetic training conditions with feedback (but not without feedback) can explain aspects of adult phonetic learning under laboratory conditions, and, perhaps, aspects of first language learning under natural settings. These findings and suggestions on the role of feedback in adult non-native phonetic training and, perhaps, in first language acquisition, can be better understood within Seitz and Watanabe's perceptual learning model in which learning occurs due to the coincidence of stimulus driven activity and the release of nonspecifically acting reinforcement and motivational signals (Grossberg & Merrill, 1996; Seitz & Watanabe, 2005).

Both task-relevant and task-irrelevant learning contribute to our understanding of how a brain solves the *stability-plasticity dilemma*; that is, how a brain can learn quickly without catastrophic forgetting. Adaptive Resonance Theory (Carpenter, 2001; Grossberg, 1980, 2007) explains how laminar neocortical circuits enable both *inter*cortical attentive feedback from higher cortical levels and *intra*cortical pre-attentive feedback from superficial layers of the same cortical region to accomplish this goal. The intracortical feedback loops self-stabilize slow perceptual learning without attention or awareness, and thus provide a plausible mechanism for the task-irrelevant learning observed in the current study. The cooperation and competition among these distinct but interacting pre-attentive and attentive processes provides a framework for investigating the conditions under which task-irrelevant learning can occur, without being inhibited by the "biased competition" properties of focused attention. For example, task-irrelevant learning may fail to engage the ART mismatch-reset process. New category learning could occur due to an interaction of filtering, competition, and intracortical resonance without the benefit of mismatch-mediated search and vigilance control processes. Further research will be required to verify this prediction.

An important question is whether task-irrelevant perceptual learning would benefit Japanese listeners. While our results cannot support strong claims regarding this issue, it should be noted that six of the TIPL group participants were native Japanese speakers. We found that five of them showed learning for the paired formant transition, with an average improvement of 23.2 ± 9.5%, while one showed a 23.1% degradation of performance. These results suggest that the procedure is effective for native Japanese speakers as it is for native English speakers. However, the threshold improvements for both native English and Japanese speakers in both training paradigms (implicit and explicit) were specific to formant transitions that were trained. Specificity of learning is typical of studies of perceptual learning, which have shown that performance improvements often do not transfer to untrained stimuli (for review see Fahle (2004)). These specificity effects are often considered as evidence that the learning effects result from changes in the sensory representation of the trained stimuli (Ahissar & Hochstein, 2004; Fahle, 2004) or in the read-out from the sensory areas (Dosher & Lu, 1998). The observation of specificity in our study is also in line with the aforementioned requirement for variability in training tokens in order for learning to generalize to untrained conditions (speakers, context, etc.) and suggests that L2 phonetic learning may be more akin to general perceptual learning. Along these lines, the fact that learning of the formant transitions is specific to frequency and sweep direction suggests that for TIPL to be effective for L2 phonetic learning, a range of simple distinctions from the new language's phonetic repertoire, such as both upward and downward formant transitions at a range of relevant frequencies, would need to be trained. Given this, further

research will be required before clear benefits to L2 learners can be achieved.

The fact that the magnitude of learning was similar between the implicit and adaptive trainings is at first glance surprising. While it is conceivable that learning in the two paradigms is due to independent mechanisms (Poldrack et al., 2001), an alternative explanation is that much learning achieved through explicit training occurs implicitly. That is, learning from the adaptive task is due to subjects' performance yielding an appropriate schedule of reinforcement to the learned stimuli, not simply due to their explicit attention towards the stimuli. Recent research of visual learning shows that attention to stimuli can actually hinder perceptual learning (Gutnisky et al., 2009; Tsushima, Seitz, & Watanabe, 2008). This finding that benefits in the perception of formant transitions achieved through task-irrelevant learning are as large as in explicit adaptive training may imply that learning of speech sounds may generally be achieved through implicit learning mechanisms.

This view of learning as emerging unintentionally, through task-irrelevant use, within a rich phonetic, linguistic and social context, although ecologically appealing, until now, has not been explored experimentally. For example, traditional training paradigms for adult second language learners in non-native phonetic contrasts use training procedures with explicit category labels and feedback after each stimulus presentation (Lively et al., 1993; Logan, Lively, & Pisoni, 1991). Obviously this is not how infants and adults learn new sounds. An interesting exception to traditional training techniques is a recent study by Wade and Holt (2005). Subjects acquired unintentionally new auditory categories after playing a 30-min "game," during which different nonspeech tokens of distinct sound categories were systematically correlated with the appearance of discrete targets. Such training paradigms have important implications for our understanding of how sensory learning is achieved by the brain and can inform how best to teach people to improve their sensory skills.

Acknowledgements

AS and TW supported by NIH (R01 EY015980-04A2, R21 EY017737-02) and NSF (BCS-0549036), and SG, YT and TW by CELEST, an NSF Science of Learning Center (SBE-0354378). We thank Erin M. Ingvalson and Lori L. Holt of Carnegie Mellon University for providing formant specifications for synthesizing r/l based on Yamada and Tohkura (1990) and for useful discussion. We also thank Daniel Khafi and Shao-Chin Hung for help conducting the experiments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cognition.2010.03.004.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America, 90*(12), 5718–5722.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457–464.

Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *Journal of the Acoustical Society of America, 119*(3), 1616–1625.

Bradlow, A. R. (2008). Training non-native language sound patterns. In J. Hansen & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). Philadelphia, PA: John Benjamins.

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.

Carpenter, G. A. (2001). Neural-network models of learning and memory: Leading questions and an emerging framework. *Trends in Cognitive Sciences, 5*, 114–118.

Choi, H., Seitz, A. R., & Watanabe, T. (2009). When attention interrupts learning: Inhibitory effects of attention on TIPL. *Vision Research, 49*(21), 2586–2590.

Dosher, B. A., & Lu, Z. L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences of the United States of America, 95*(23), 13988–13993.

Dranias, M. R., Grossberg, S., & Bullock, D. (2008). Dopaminergic and non-dopaminergic value systems in conditioning and outcome-specific revaluation. *Brain Research, 1238*, 239–287.

Fahle, M. (2004). Perceptual learning: A case for early selection. *Journal of Vision, 4*(10), 879–890.

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

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America, 100*(13), 8030–8035.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*(6), 509–516.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (pp. 233–374). New York: Academic Press.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1–51.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*, 23–63.

Grossberg, S. (2007). Consciousness CLEARS the mind. *Neural Networks, 20*, 1040–1053.

Grossberg, S., & Merrill, J. W. L. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience, 8*, 257–277.

Gutnisky, D. A., Hansen, B. J., Iliescu, B. F., & Dragoi, V. (2009). Attention alters visual plasticity during exposure-based learning. *Current Biology*.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, Mass: MIT Press.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 363*(1493), 979–1000.

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America, 100*(15), 9096–9101.

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*(3 Pt 1), 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.

Loebach, J. L., & Pisoni, D. B. (2008). Perceptual learning of spectrally degraded speech and environmental sounds. *Journal of the Acoustical Society of America, 123*(2), 1126–1139.

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*(2), 874–886.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–111.

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: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2*(2), 89–108.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Dalattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, r, l/ in English. *Word, 13*, 25–43.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognition Science, 10*(5), 233–238.

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words*. Cambridge: MIT Press.

Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., et al. (2001). Interactive memory systems in the human brain. *Nature, 414*(6863), 546–550.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Saffran, J. R., Werker, J., & Werner, L. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In R. Siegler & D. Kuhn (Eds.), *Handbook of child development* (pp. 58–108). New York: Wiley.

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognition Science, 9*(7), 329–334.

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron, 61*(5), 700–707.

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research, 49*(21), 2604–2610.

Stevens, K. (1998). *Acoustic phonetics*. Cambridge: MIT Press.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503–2522.

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience, 18*, 1029–1043.

Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. *Current Biology, 18*(12), R516–517.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, & Behavioral Neuroscience, 7*(1), 53–73.

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America, 118*(4), 2618–2633.

Yamada, R. A., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. In *Proceedings of the 1st international conference on spoken language production* (pp. 757–760). Japan: Kobe.