COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 1390–1405
Copyright © 2011 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01192.x

Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization

Sung-joo Lim, Lori L. Holt

Department of Psychology, Carnegie Mellon University

Received 9 June 2010; received in revised form 23 February 2011; accepted 25 February 2011

Abstract

Although speech categories are defined by multiple acoustic dimensions, some are perceptually weighted more than others and there are residual effects of native-language weightings in non-native speech perception. Recent research on nonlinguistic sound category learning suggests that the distribution characteristics of experienced sounds influence perceptual cue weights: Increasing variability across a dimension leads listeners to rely upon it less in subsequent category learning (Holt & Lotto, 2006). The present experiment investigated the implications of this among native Japanese learning English /r/-/l/ categories. Training was accomplished using a videogame paradigm that emphasizes associations among sound categories, visual information, and players' responses to videogame characters rather than overt categorization or explicit feedback. Subjects who played the game for 2.5 h across 5 days exhibited improvements in /r/-/l/ perception on par with 2–4 weeks of explicit categorization training in previous research and exhibited a shift toward more native-like perceptual cue weights.

Keywords: Learning; Non-native speech categorization; Speech perception; Categorization; Videogame training; Adult plasticity; Auditory learning; Second language learning

1. Introduction

Before we are native speakers, we are native listeners. In the first year of life, experience with the native language begins to shape how we perceive speech (Werker & Tees, 1984). Although this learning is well underway before infants speak their first words, speech category learning and refinement has a lengthy developmental course such that even 12-year-olds are not entirely adult-like (Hazan & Barrett, 2000; Nittrouer, 2004).

Correspondence should be sent to Sung-joo Lim, Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: sungjol@andrew.cmu.edu

By adulthood, however, we are experts in speech categorization. The perceptual space is ''warped'' (Kuhl et al., 2008) such that it reflects regularities of native speech input, exaggerating between-category differences (promoting discrimination), and diminishing within-category differences (promoting generalization, Näätänen, 2001). Moreover, although speech sounds are characterized by multiple co-varying acoustic dimensions (Lisker, 1986), they are not equally informative in differentiating speech categories, and adult listeners give greater perceptual weight to acoustic dimensions that reliably differentiate native speech categories (Nittrouer, 2004; Yamada & Tohkura, 1990). An example is English /r/-/l/ (as in *rock* and *lock*). Whereas the onset frequencies of the second (F2) and the third (F3) formants each co-vary with native-English speakers' /r/-/l/ utterances, F3 onset frequency is a much more reliable dimension for distinguishing English speakers' /r/ and /l/ productions (Lotto, Sato, & Diehl, 2004). Accordingly, native-English listeners exploit F3 onset frequency more than F2 onset frequency in categorizing /r/ and /l/ (Yamada & Tohkura, 1990). Learned perceptual representations of native speech sounds emphasize, or perceptually weight, acoustic dimensions contributing to phonological contrasts within the native language.

This expertise has consequences. Having been shaped by native-language sound structure, adults are affected by an acute and persistent influence of native speech categories in perceiving non-native speech sounds (Best, 1995; Flege, 1995). Native-Japanese adults, for example, have great difficulty distinguishing English /r/-/l/, categories not native to Japanese (Goto, 1971; Iverson et al., 2003; Miyawaki et al., 1975). For native-Japanese listeners, experience with the single Japanese speech category that occupies a similar perceptual space is thought to affect English categorization (Flege, 1995). Whereas F3 onset frequency distinguishes /r/-/l/ in native-English speech production and perception (Lotto et al., 2004; Yamada & Tohkura, 1990), most native-Japanese listeners appear to rely instead on the less-informative F2 onset frequency for English /r/-/l/ categorization (Iverson et al., 2003; Lotto et al., 2004; Yamada & Tohkura, 1990). Presumably, this arises from expertise with Japanese. It is not a consequence of a lack of psychoacoustic *sensitivity* to the F3 acoustic dimension because adult Japanese listeners rely upon F3 to distinguish /d/-/g/ (sounds with Japanese counterparts; Mann, 1986), they exhibit compensatory adjustment in perception of a syllable that follows /r/-/l/ sounds differing only in F3 (Mann, 1986) and they are highly accurate in discriminating nonspeech chirps created by stripping /r/-/l/ sounds of all acoustic information except F3 (Miyawaki et al., 1975). Japanese listeners' difficulty in relying on the F3 for /r/-/l/ categorization persists even with explicit training (Iverson, Hazan, & Bannister, 2005) and after immersion in English (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Gordon, Keyes, & Yung, 2001).

Laboratory-based speech training studies provide evidence of some plasticity in the adult system to support non-native category learning, although the system is clearly not as flexible as in earlier development (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Goudbeek, Cutler, & Smits, 2008; Iverson et al., 2005; Jamieson & Morosan, 1989; Logan, Lively, & Pisoni, 1991; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; Pisoni, Logan, & Lively, 1994). Recent research has underscored the importance of acoustic variability in training; including multiple speakers and phonetic contexts seems to aid learning and generalization (Bradlow et al., 1997; Iverson et al., 2005; Jamieson & Morosan, 1989; Lively, Logan, & Pisoni, 1993; McCandliss et al., 2002). In such studies, participants tend to improve in their ability to reliably categorize non-native speech over the course of training. However, these studies have relied upon extensive training across 2–4 weeks (Bradlow et al., 1997; Lively et al., 1993; Logan et al., 1991), explicit trial-by-trial performance feedback during training is typical, and the nature of the influence of high variability upon learning is, as yet, quite poorly understood with regard to its influence on aspects of adult categorization like perceptual cue weighting (although see Iverson et al., 2005). There remain important gaps in our models of speech category learning.

In parallel, researchers have begun to investigate general principles of auditory category learning, with attention to the specific challenges posed by speech category learning (Goudbeek, Swingley, & Smits, 2009; Guenther, Husain, Cohen, & Shinn-Cunningham, 1999; Holt & Lotto, 2006; Holt, Lotto, & Diehl, 2004; Mirman, Holt, & McClelland, 2004; Wade & Holt, 2005). In these studies, the use of novel nonspeech sounds affords a degree of control over input that is impossible with speech categories, providing an opportunity to explicitly manipulate experience. For example, in training listeners to categorize novel nonspeech sounds, Holt and Lotto (2006) demonstrated that listeners' perceptual cue weight for an acoustic dimension can be decreased by introducing variability along the dimension; listeners are induced to make use of less-variable acoustic dimensions in categorization. This demonstrates that listeners' sound categorization is not driven solely by the acoustic signature of a particular sound but, also, by the relationship of the sound to the distribution of previously experienced sounds.

This has implications for adult non-native speech category learning because non-native perceptual cue weighting may be less-than-optimal for categorization (as exemplified by Japanese listeners' reliance on the less-informative F2 onset-frequency dimension for English /r/-/l/). Although distributional information and, in particular, acoustic variability, has been widely suggested to aid in training non-native speech categories (Bradlow et al., 1997; Lively et al., 1993; Logan et al., 1991), there is as yet no detailed understanding of how this affects learning. In the present study, we hypothesize from Holt and Lotto's (2006) finding with nonspeech auditory categorization that training native-Japanese adults with distributions of speech exemplars sampled such that the less informative, but perceptually ''preferred'' dimension (F2 onset frequency) is sampled with high variability whereas the more-informative, but less perceptually weighted dimension (F3 onset) is sampled with relatively less variability will result in more native-like perceptual cue weighting with greater reliance on F3 onset following training.

We also aim to investigate whether learning may occur without trial-by-trial performance-based feedback that is typical of studies of adult non-native speech category learning, but less-than-typical of natural speech-learning environments. Several previous studies have investigated adult speech category learning with unsupervised learning paradigms that do not involve performance feedback (Goudbeek et al., 2008; McCandliss et al., 2002; McClelland, Fiez, & McCandliss, 2002). Where direct comparisons have been made for adult speech category learning, supervised learning with trial-by-trial feedback exceeds unsupervised learning without performance feedback (Goudbeek et al., 2008; McCandliss et al., 2002; Vallabha & McClelland, 2007) but learning can be observed in unsupervised paradigms. In some ways, however, the approach of eliminating performance feedback while maintaining an explicit categorization training task is similarly unnatural. It is possible that making explicit category judgments (with or without feedback) may not draw upon the same processes by which humans learn to categorize natural sounds, including speech.

Bridging this gap in a study investigating nonspeech auditory category learning, Wade and Holt (2005) developed an active task requiring integration of multimodal information, but involving no explicit categorization responses or categorization-performance feedback. In the study, participants navigated through a space-invaders-style videogame involving visual aliens, each associated with a category of multiple nonspeech sound exemplars. As in the natural environment, sound plays a functional role in the game and it is richly correlated with other perceptual and motor information. Wade and Holt (2005) report nonlinguistic sound category learning and generalization after just 30 min of game play.

Note, however, that although listeners do not engage in an explicit categorization task and there is no feedback directly associated with sound categorization, the videogame paradigm is not completely ''unsupervised'' in the manner of studies of adult non-native speech category learning (Goudbeek et al., 2008; McCandliss et al., 2002); feedback arrives in the form of succeeding or failing in the mission of the game. As an indication that this intermediate form of supervision supports learning, Wade and Holt (2005) found that the nonspeech auditory category learning induced by training within the game exceeded that of another group of participants who explicitly attempted to learn the categories in an unsupervised manner, without feedback, for nonlinguistic sounds that mimic the complexity of speech categories. We exploit the videogame method here to investigate perceptual cue weighting for non-native sounds among adults training on non-native speech categories.

Native-Japanese adults played a version of the Wade and Holt (2005) videogame in which four aliens were associated with four English speech categories, /ra/, /la/, /da/, and /ga/. Because the /da/-/ga/ categories are similar to native-Japanese speech categories, we expected performance be high and relatively unchanged by training except inasmuch as participants adapted to the speech training stimuli during pretest. Thus, they served as controls. The /ra/ and /la/ categories were the primary test; categories were sampled with high variability along the F2 dimension to test whether training would shift Japanese listeners' perceptual cue weights toward F3, as predicted by the nonspeech auditory category learning results of Holt and Lotto (2006). Finally, a control group of native-Japanese adults played the game with the artificial nonspeech sound categories used by Wade and Holt (2005) to assure that experience in the game itself, independent of the sounds experienced, did not contribute to posttraining speech performance. To assess learning, all participants completed a battery of speech categorization tests before and after playing the game across five consecutive daily sessions.

2. Method

Twenty-seven native-Japanese adults (ages 24–40) living in Pittsburgh were recruited. They had lived in English-speaking countries for fewer than 2.5 years (*M* = 8.8 months) and had begun learning English in junior high school (*M* = 12.1 years). Thirteen of these participants were randomly assigned to the Training condition; the rest served as a Control.

Testing took place in sound-isolating booths with E-prime-controlled presentation and response collection for pre and posttests (Schneider, Eschman, & Zuccolotto, 2002). A custom-designed videogame was used for training, identical to Wade and Holt (2005) except in the sound stimuli defining categories. Participants used a standard computer keyboard to register responses and control the game. A video monitor placed directly in front of participants provided the visual stimuli and written instructions. There was no performance feedback regarding pre and posttest responses. Synthesized speech stimuli were created using a speech synthesizer (Klatt & Klatt, 1990) to afford detailed control over F2 and F3 onset frequencies. All stimuli were sampled at 22 kHz, RMS-matched in amplitude, and presented at approximately 70 dB.

On Day 1, there were four pretests: 1) *Categorization:* Listeners responded to 250-ms synthesized speech syllables as /ra/, /la/, /da/, and /ga/ with no feedback. For each category, participants responded to eight training exemplars and 13 novel, untrained category exemplars to test generalization. Orthogonal manipulation of the second (F2) and third (F3) onset formant frequencies defined acoustic stimulus space (see Fig. 1); other parameters were held constant (see Table 1; synthesis parameters followed Yamada & Tohkura, 1990 and Lotto & Kluender, 1998). Category identity of these synthesized sounds was verified with nine native-English listeners who performed at ceiling, *M* = 90.4%. The /da/-/ga/

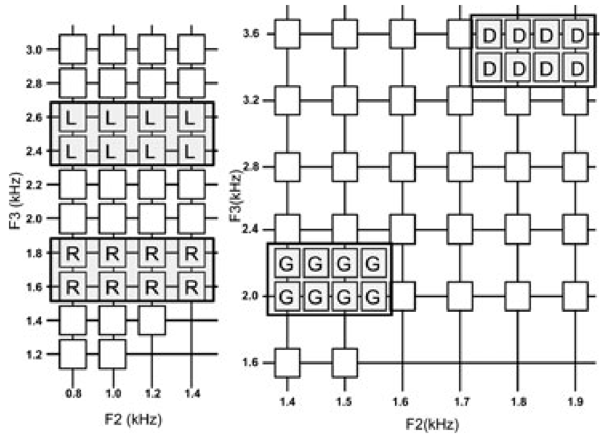

Fig. 1. Schematic diagram of training stimuli. The left panel shows the F2 and F3 onset-frequency structure of /r/-/l/ stimuli, whereas the right panel plots /d/-/g/ stimuli. All stimuli on the grids were used during perceptual cue weighting tasks. The stimuli highlighted with shaded rectangles were used during the training and categorization testing. Additional interleaved stimuli within the shaded region (not pictured) served as generalization stimuli.

Table 1
Parameters for synthesizing training stimuli using Klatt and Klatt (1990)

	Duration (ms)	F1	F2	F3	F4	F5
/da/-/ga/ synthesis parameters						
Consonant onset frequency	0	400	*	*	3400	3950
Vowel/a/onset	75	750	1220	2465	3400	3950
	245	750	1220	2465	3400	3950
/ra/-/la/ synthesis parameters						
Consonant onset frequency	0	400	*	*	3400	3950
	80	400	*	*	3400	3950
Vowel/a/onset	150	750	1220	2465	3400	3950
	245	750	1220	2465	3400	3950

Note. The ''*'' denotes the onset-frequency manipulations of F2 and F3 as shown in Fig. 1. Vowel/a/onset parameters are equivalent for all stimuli.

stimuli served as a baseline measure of English categorization performance among Japanese listeners. Each stimulus was categorized 10 times.

2–3) *Perceptual cue weighting:* Separate categorization tests were administered to assess perceptual cue weighting of F2 versus F3 onset frequency for English /r/-/l/ and /d/-/g/. For /r/-/l/, participants categorized 10 repetitions of 37 stimuli sampling a two-dimensional grid with orthogonal variation along the second (F2) and third (F3) onset formant frequency dimensions (after Yamada & Tohkura, 1990); all other acoustic parameters were held constant (see Table 1). This F2 × F3 manipulation produced stimuli that varied perceptually from *right* to *light* (after Yamada & Tohkura, 1990; see Fig. 1). The purpose of this task was to measure listeners' relative reliance on F2 versus F3 in categorizing English /r/-/l/. A similar test of 32 variations of /da/-/ga/ stimuli in F2 × F3 (see Fig. 1) onset-frequency space was administered separately as a control. All stimuli were synthesized using the Klatt and Klatt (1990) synthesizer to strictly manipulate F2 and F3 while holding all other acoustic dimensions constant. Participants received no feedback.

4) *Generalization to natural speech:* To assess generalization, listeners identified 224 naturally spoken /r/-/l/ minimal pairs (e.g., *road-load*) uttered by three (two male) talkers. These stimuli were drawn from previous research (Lively et al., 1993; Logan et al., 1991) investigating the generalization of speech category training to natural native-English utterances. There were 32 filler items possessing other English consonants more easily categorized by Japanese listeners; these served as a check for task compliance.

2.1. Videogame training

Following the Day 1 pretests, an experimenter instructed participants how to play the videogame before 15 min of play. On Days 2–4, participants played the game for 30 min. On Day 5, participants played the game 20 min before completing posttests.

Training was accomplished using the Wade and Holt (2005) videogame paradigm. Participants navigated a pseudo-3D tunnel using a computer keyboard, encountering four alien creatures, each with a unique shape, color, and movement pattern. Each alien originated from a particular quadrant of the virtual environment (with a jitter of random noise to somewhat alter position). Participants' task was to capture two ''friendly'' aliens and destroy two ''enemy'' aliens; identity as friends or enemies was conveyed via the shape and color of a shooting mechanism on the screen. Thus, there was a great deal of visual and spatial information with which to succeed in the task, independent of sound categories, and early in the game aliens moved slowly providing participants time to experience these rich and consistent regularities.

Each alien also was assigned a sound category composed of eight unique exemplars. Each time an alien appeared, a randomly drawn exemplar from its category was presented repeatedly through the duration of the alien's appearance. Consequently, the auditory category defined by multiple, acoustically variable exemplars co-occurred over the course of the game with the visual image of the alien, the spatial quadrant from which the alien originated, and the distinctive motor/tactile patterns involved in capturing or destroying it. Whereas these regularities are simplistic compared to the rich set of natural multimodal cues, it presents much richer contextual support than is typically present in supervised learning paradigms, where categories co-occur with their labels and feedback assignments or in unsupervised laboratory learning paradigms where these relationships do not exist.

In the game, exposure to patterns of co-variance is fairly incidental; knowledge of the acoustic exemplars is of no apparent consequence to performance and participants are instructed only to play the game. Aliens initially appear near the center of the screen and approach slowly. However, the game becomes progressively more difficult such that aliens' approach speeds up gradually. Moreover, characters begin approaching from locations that are increasingly peripheral to the center of the screen. These factors require players to react more quickly; as aliens' origins become still more distal, targeting becomes nearly impossible without quick categorization of sound patterns (which predicts the alien and its quadrant of origin). Good performance at the highest levels requires a repeated, instantaneous, functionally oriented use of sound categories that is generally not demanded of participants in supervised or unsupervised auditory category training studies.

For fourteen (Control) participants, these sound categories were composed of the complex nonspeech stimuli used by Wade and Holt (2005). Control participants completed the speech pre/posttests, but they did not have training with English speech categories, protecting against the possibility that experience with the videogame supports non-native category learning indirectly. The remaining (Trained) participants experienced synthesized speech syllables drawn from English /ra/, /la/, /da/, and /ga/ categories, each defined by eight unique exemplars sampling the F2 × F3 acoustic space. Following the prediction that high variability along the less-diagnostic (F2) cue and low variability along the diagnostic (F3) cue would shift native-Japanese listeners' /r/-/l/ to more native-like perceptual cue weighting (Holt & Lotto, 2006), training stimuli were synthesized to strictly manipulate these dimensions while holding other acoustic information constant (see Table 1).

Posttests: After playing the game 20 min on Day 5, the four perception pretests were repeated to assess posttraining changes in performance. Test order was counterbalanced across participants and between pre/posttest sessions except that the Category Learning test was administered last on Day 1 and first on Day 5 to protect against any short-term influences of the training stimulus distributions extending into the Perceptual Cue Weighting test.

3. Results

3.1. Category learning

A *t* test revealed that native-Japanese participants' English pretest /r/-/l/ categorization accuracy was very poor compared to native-English listeners (95.5%, $t(34) = 12.87$, $p \leq .0005$), but significantly above chance (Trained: 56.1%, $t(12) = 2.62$, $p = .02$; Control: 56.0%, $t(13) = 2.28$, $p = .04$). In comparison, native-Japanese participants' pretest /da/-/ga/ categorization was more accurate (91%, $t(26) = 16.03$, $p \leq .0005$) and indistinguishable from native-English listeners' performance, $t(34) = 1.80$, $p = .08$, verifying /da/-/ga/ as a baseline for accurate native-Japanese categorization.

A repeated-measures analysis of variance revealed a significant interaction of Condition (Training, Control) and Test (pre, post) for native-Japanese listeners' English /ra/-/la/ categorization, $F(1, 25) = 8.65$, $p = .007$, $\eta_p^2 = 0.26$. Trained listeners improved 18.5% in /ra/-/la/ categorization from pre to posttest, $t(12) = 4.22$, $p = .001$, whereas Control listeners improved 4.7%, $t(13) = 2.43$, $p = .031$ (see Fig. 2). The significant improvement observed for Control participants likely reflects a small, consistent benefit of experiencing the synthesized speech stimuli during pretest that facilitates posttest categorization. This interpretation is supported by improvements by Control ($M = 5.3\%$, $t(13) = 4.39$, $p = .001$) and Trained ($M = 6.0\%$, $t(12) = 3.74$, $p = .003$) listeners in /da/-/ga/ categorization of about the same magnitude. Thus, about 5% of observed improvements are likely to be the consequence of tuning speech perception to the specific acoustic features of the synthesized speech.

With a 18.5% improvement, the Training group's English /ra/-/la/ categorization well exceeded these levels with just 2.5 h of videogame play, $F(1,25) = 7.27$; $p = .012$; $\eta_p^2 = 0.225$. To put this learning in context, standard laboratory training with explicit-feedback about overt categorization of English /r/-/l/ yields comparable learning effects across 2–4 weeks (about 45 h) of training for native-Japanese listeners with English proficiency similar to the current participants (Bradlow et al., 1997). Of particular note, the present improvement in /r/-/l/ categorization resulted from incidental categorization training without explicit performance feedback or overt categorization. Immersive experience within an environment possessing richly correlated perceptual cues associated with the speech categories appears to be very effective in promoting speech category learning, even without feedback.

3.2. Perceptual cue weight

Following previous research (Holt & Lotto, 2006), the correlation of F2 and F3 onset frequencies with categorization responses was calculated for each listener and normalized to 100% as a means of quantifying the relative perceptual cue weight of the two acoustic

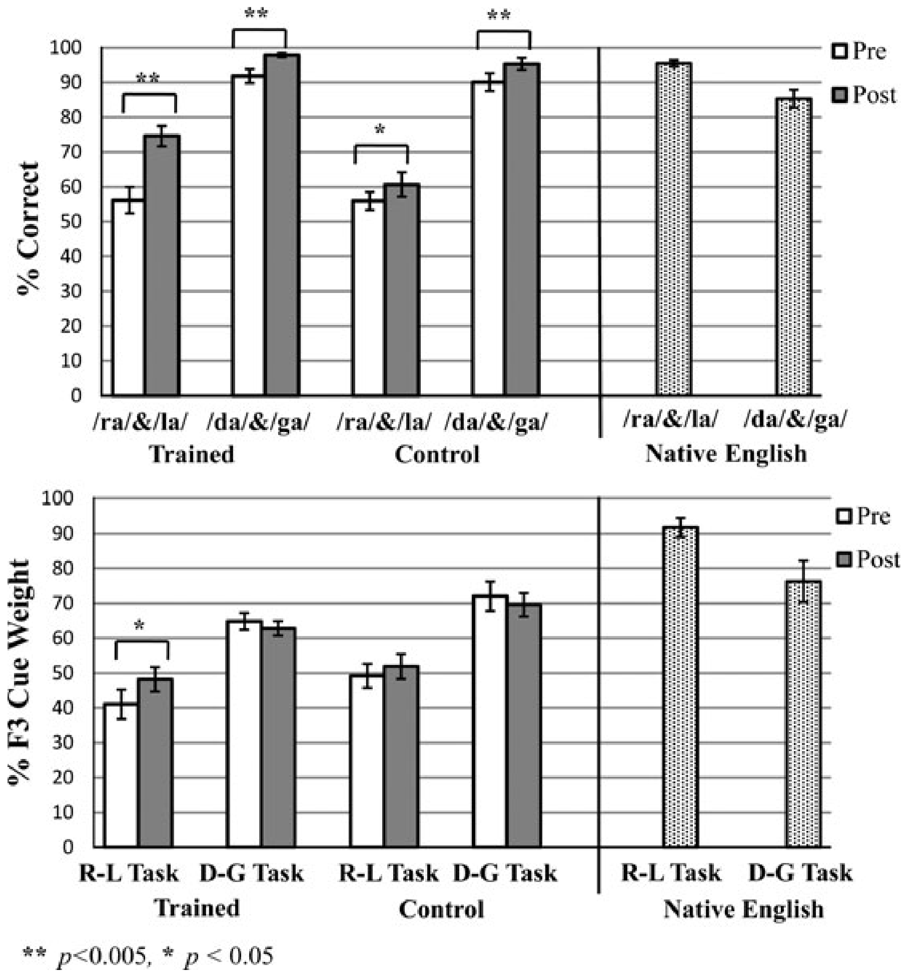

Fig. 2. Results of videogame training. The top panel illustrates pre versus posttest percent correct categorization of /r/-/l/ and /d/-/g/ for Trained versus Control participants in comparison to native-English listeners. The bottom panel shows the change in perceptual weighting of F3 in categorization responses from pre to posttest among Trained and Control listeners on /r/-/l/ and /d/-/g/ tasks in comparison to native-English listeners.

dimensions. Based on previous results (Holt & Lotto, 2006), we predicted that sampling the /r/-/l/ F2 × F3 onset-frequency space with higher F2-onset-frequency variability and lower F3-onset-frequency variability would promote greater perceptual weighting of F3 among Trained native-Japanese listeners. Although the Trained group did not reach native-like levels of reliance on F3 in categorizing /r/-/l/, $t(20) = 9.12, p \leq .0005$, the short-term training did have a significant effect on perceptual cue weighting. Native-Japanese participants relied more on F3 at posttest relative to pretest ($M = 7.1\%$, $t(12) = 2.78, p = .017$). This pattern was not observed among Control participants categorizing /r/-/l/ ($M = 2.7\%$, $t(13) = 1.88, p = .082$) or for /d/-/g/ categorization for either group as shown in Fig. 2 (Trained: $t(12) = 0.69, p = .51$; Control: $t(13) = 0.96, p = .35$).

3.3. Generalization to natural speech

Training in the videogame appears to have shifted native-Japanese listeners' categorization away from the less-diagnostic F2 onset frequency and toward F3 onset frequency. Since F3 is the single acoustic dimension most diagnostic of /r/-/l/ category membership in native-English talkers' speech (Lotto et al., 2004), this may improve categorization of natural speech even without experience with natural speech during training. In modest support of this hypothesis, Trained listeners exhibited a trend in improved recognition of naturally spoken /r/-/l/ words ($M = 8.5\%$, $t(12) = 1.96$, $p = .074$), whereas the Control group showed 5.4% improvement ($t(13) = 1.49$, $p = .159$). The fact that listeners' performance was already above chance (50%) at pretest ($M = 67.8\%$, $t(26) = 8.65$, $p \leq .0005$), in addition to individual differences in performance, may have resulted a lack of significant improvement from the training. However, the trend suggests that the learning resulting from just 2.5 h of incidental training with stylized synthetic speech in a videogame may have implications for natural spoken word recognition.

4. Discussion

Infants' speech perception rapidly tunes to the regularities of the native language (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker & Tees, 1984) and, perhaps as a consequence, it is notoriously difficult for adults to learn non-native speech categories that assimilate to native-language categories (Best, 1995; Flege, 1995). As such, investigations of adults learning non-native speech categories have provided an opportunity to examine the limits of adult plasticity and perceptual learning and to search for effective methods to facilitate adult speech category acquisition via directed short-term training.

Previous studies have stressed the importance of high acoustic variability in promoting non-native speech category learning (Iverson et al., 2005; Lively et al., 1993; Logan et al., 1991). However, although there is general agreement that the distributional characteristics of experienced speech contribute to the effects of variability on learning, the details are not yet clear. Here, significant learning was evoked in a short time without high-variability training; training tokens were created with a single synthesized ''voice'' and varied minimally in acoustic space. Perhaps important, however, was the acoustic variability that did exist in the present training stimuli.

Motivated by findings from how adults learn nonlinguistic sound categories (Holt & Lotto, 2006), the current study investigated whether high variability along a perceptually preferred, but less reliable, acoustic dimension (F2 onset) would shift non-native listeners' perceptual cue weights toward more reliable information (F3 onset), thus promoting more native-like categorization. The present results parallel those observed for nonlinguistic sound category learning (Holt & Lotto, 2006); following training non-native categorization accuracy was improved and listeners relied more on the less-variable (F3 onset) acoustic dimension. The approach of enhancing variability along a preferred, but poorly diagnostic acoustic dimension promoted shifts in perceptual cue weighting toward more native-like patterns.

It is notable that the observed learning was evoked without explicit categorization training or performance-based feedback. Participants simply attempted to navigate the videogame, in the process experiencing a rich set of multimodal regularities associated with the speech sound categories. Nonetheless, just 2.5 h of training resulted in adult non-native speech category learning comparable to learning observed from 2 to 4 weeks of training with standard categorization training paradigms with explicit-feedback (Bradlow et al., 1997; Iverson et al., 2005; Lively et al., 1993; Logan et al., 1991). This suggests the possibility that functional use of sound may facilitate complex, multidimensional category learning even without an overt categorization task. An open question is whether native-Japanese listeners might achieve even greater English speech category learning with longer training periods training within the videogame paradigm. Of potential relevance to this issue, participants playing the videogame across 5 days achieved higher nonspeech sound categorization accuracy (Liu & Holt, 2011) than did participants who played for just 30 min (Wade & Holt, 2005). The learning reported here may not represent upper bounds of adult plasticity for language learning. The details of the learning mechanisms evoked by videogame training are as yet unclear, but evidence from other research domains is informative.

Previous research on adult non-native speech category learning highlights that although learning is more robust in paradigms with performance feedback, it *can* be observed without feedback (McCandliss et al., 2002). This is consistent with research that demonstrates learning can be driven with extrinsic rewards like performance feedback or monetary gains (Seitz, Kim, & Watanabe, 2009; Tricomi, Delgado, McCandliss, McClelland, & Fiez, 2006) or by intrinsically generated performance evaluation in lieu of feedback (Seitz & Watanabe, 2009). As highlighted in the introduction, the present videogame paradigm is unique in that it lies between these two endpoints. Participants do not engage in explicit categorization and do not received performance feedback about categorization. Yet learning is not entirely passive or unsupervised; feedback arrives in the form of success or failure in achieving one's goals in the game and there are multiple, correlated multimodal events and objects that covary with speech category membership. These characteristics may engage intrinsically generated learning signals to a greater extent than passive, unsupervised training paradigms and, perhaps, even to a greater degree than extrinsic performance feedback. Supportive of the first possibility, Wade and Holt (2005) found that the nonspeech auditory category learning within the game exceeded unsupervised learning of the same sounds. Supportive of the second possibility, direct attention to target stimuli can sometimes actually hamper perceptual learning (Gutnisky, Hansen, Illiescu, & Dragoi, 2009; Tsushima, Seitz, & Watanabe, 2008).

Videogames produce robust perceptual learning (e.g., Green & Bavelier, 2007; Li, Polat, Makous, & Bavelier, 2009) and may be highly effective at activating the striatal reward system of the brain (e.g., Koepp et al., 1998), providing the intrinsic learning signals. Based on human neuroimaging research, Tricomi et al. (2006) propose that tasks that include goal-directed action for which there is a positive or negative outcome contingent on one's behavior, tasks in which actions are performed in the context of expectations about outcomes, and tasks in which individuals have incentive to perform well (Delgado, Stenger, & Fiez, 2004) are most likely to robustly activate striatal reward system processing (within the caudate nuclei, in particular). Recruitment of the striatal reward system may facilitate learning through feedback to perceptual representations and, additionally, may affect learning through its influence on other mechanisms, such as Hebbian learning (Vallabha & McClelland, 2007), by serving as an informative signal to guide learners to better differentiate available information (Callan et al., 2003; McCandliss et al., 2002), as for example F2 versus F3 onset frequencies in the present task. The videogame paradigm fits these optimal task characteristics quite well. The relatively high motivation and engagement elicited by videogames (especially compared to standard, overt laboratory categorization tasks) may evoke greater intrinsic reward-based processing supportive of learning. The intrinsic reward of success in the game, of accurately predicting and acting upon upcoming events, may be a powerful signal to drive learning.

To draw this back to natural language learning, the striatal reward system responds to many reinforcers. Statistical language learning mechanisms (see Kuhl, 2004 for review), typically studied in wholly passive exposure learning paradigms in the laboratory (e.g., Saffran, Aslin, & Newport, 1996), may be supported in natural language learning by modulation from attentional and motivational factors (Kuhl, Tsao, & Liu, 2003) and contingent extrinsic reinforcers like social cues (Goldstein, King, & West, 2003). The intermediate nature of the videogame training paradigm (active, not passive; lacking performance feedback without being entirely unsupervised) thus may better model some aspects of language learning than wholly passive exposure paradigms or paradigms with overt categorization and explicit performance feedback.

The rich correlations of the to-be-learned non-native speech categories with other perceptual and motor cues in the videogame also may have supported learning. Temporal coincidence of cues, irrespective of their relationship to a task, can generate intrinsic reinforcement signals (Seitz & Dinse, 2007; Seitz & Watanabe, 2005; Seitz et al., 2010) and performance feedback from typical training paradigms generates a learning signal only *after* perception and decision processes, thereby eliminating potentially beneficial signals originating from temporal coincidence. The presence of correlated visual cues with an artificial auditory language facilitates passive statistical learning of sequential probabilities in adults and infants compared to unimodal auditory presentation (Thiessen, 2010). Moreover, co-occurrence of an additional cue, such as social gaze, improves infants' learning of visual objects in complex learning environments (Wu, Gopnik, Richardson, & Kirkham, 2010).

Many accounts presume that distributional learning drives speech category learning, but it is not yet well understood to which distribution statistics listeners are sensitive, how feedback of various forms may influence distributional learning, how acoustic dimensions are perceptually weighted, and how task affects the warping of perceptual space. Carefully manipulating experience with non-native speech categories within a more natural learning environment like the present videogame provides an opportunity to investigate these issues in greater depth to discover constraints on auditory learning and plasticity relevant to speech categorization. Moreover, the present results highlight how studies of general auditory category learning of nonlinguistic sounds can lead to detailed, testable predictions about speech category learning. This emphasizes the domain generality of the learning and thus argues that studies of adult plasticity for language learning have implications for understanding the behavioral and neural bases of learning in other domains.

Acknowledgments

Awards to L.L.H. from the National Science Foundation (BCS-0746067) and the National Institutes of Health (R01DC004674) and training grants to S.J.L. from National Science Foundation (0549352) and National Institutes of Health (T32GM081760) supported the work.

References

Aoyama, K., Flege, J. E., Guion, S., 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.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

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.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language contrast. *NeuroImage*, *19*, 113–124.

Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004). Motivation-dependent responses in the human caudate nucleus. *Cerebral Cortex*, *14*(9), 1022–1030.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

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*, *100*, 8030–8035.

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

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds ''l'' and ''r.'' *Neuropsychologia*, *9*, 317–323.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, *50*, 109–125.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933.

Green, C. S., & Bavelier, D. (2007). Action videogame experience alters the spatial resolution of attention. *Psychological Science*, *18*(1), 88–94.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, *106*, 2900–2912.

Gutnisky, D. A., Hansen, B. J., Illiescu, B. F., & Dragoi, V. (2009). Attention alters visual plasticity during exposure-based learning. *Current Biology*, *19*(7), 555–560.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*, 377–396.

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., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of America*, *116*, 1763–1773.

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. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, *43*, 88–96.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, *87*, 820–856.

Koepp, M. J., Gunn, R. N., Lawrence, A. D., Cunningham, V. J., Dagher, A., Jones, T., Brooks, D. J., Bench, C. J., & Grasby, P. M. (1998). Evidence for striatal dopamine release during a video game. *Nature*, *393*, 266–268.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Native Language Magnet Theory Expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, *363*, 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*, *100*, 9096–9101.

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

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action videogame playing. *Nature Neuroscience*, *12*, 549–551.

Lisker, L. (1986). ''Voicing'' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11.

Liu, R., & Holt, L. L. (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. *Journal of Cognitive Neuroscience*, *23*(3), 683–698.

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.

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., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*, 602–619.

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 J. Slifka, S. Manuel, & M. Matthies (Eds.), *Proceedings of from sound to sense: 50+ years of discoveries in speech communication* (pp. 1–6). Cambridge, MA: MIT Press.

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

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology and Behavior*, *77*, 657–662.

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. *Perception and Psychophysics*, *18*, 331–340.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*, 1–21.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *Journal of the Acoustical Society of America*, *115*, 1777–1790.

Pisoni, D. B., Logan, J. S., & Lively, S. E. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In H. C. Nusbaum & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 121–166). Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Seitz, A. R., & Dinse, H. R. (2008). A common framework for perceptual learning. *Current Option in Neurobiology*, *17*(2), 148–153.

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

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (in press). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*(3), 435–443.

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

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

Thiessen, E. D. (2010) Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, *23*(6), 1093–1106.

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–R517.

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*, 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*, 2618–2633.

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.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2010). Social cues support learning about objects from statistics in infancy. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1228–1233). Austin, TX: Cognitive Science Society.

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 1990 International Conference on Spoken Language Processing* (pp. 757–760). Japan: Kobe