

Journal of Experimental Psychology: Human Perception and Performance

Discovering Functional Units in Continuous Speech

Sung-Joo Lim, Francisco Lacerda, and Lori L. Holt

Online First Publication, May 25, 2015. <http://dx.doi.org/10.1037/xhp0000067>

CITATION

Lim, S.-J., Lacerda, F., & Holt, L. L. (2015, May 25). Discovering Functional Units in Continuous Speech. *Journal of Experimental Psychology: Human Perception and Performance*. Advance online publication. <http://dx.doi.org/10.1037/xhp0000067>

Discovering Functional Units in Continuous Speech

Sung-Joo Lim
Carnegie Mellon University

Francisco Lacerda
Stockholm University

Lori L. Holt
Carnegie Mellon University

Language learning requires that listeners discover acoustically variable functional units like phonetic categories and words from an unfamiliar, continuous acoustic stream. Although many category learning studies have examined how listeners learn to generalize across the acoustic variability inherent in the signals that convey the functional units of language, these studies have tended to focus upon category learning across isolated sound exemplars. However, continuous input presents many additional learning challenges that may impact category learning. Listeners may not know the timescale of the functional unit, its relative position in the continuous input, or its relationship to other evolving input regularities. Moving laboratory-based studies of isolated category exemplars toward more natural input is important to modeling language learning, but very little is known about how listeners discover categories embedded in continuous sound. In 3 experiments, adult participants heard acoustically variable sound category instances embedded in acoustically variable and unfamiliar sound streams within a video game task. This task was inherently rich in multisensory regularities with the to-be-learned categories and likely to engage procedural learning without requiring explicit categorization, segmentation, or even attention to the sounds. After 100 min of game play, participants categorized familiar sound streams in which target words were embedded and generalized this learning to novel streams as well as isolated instances of the target words. The findings demonstrate that even without a priori knowledge, listeners can discover input regularities that have the best predictive control over the environment for both non-native speech and nonspeech signals, emphasizing the generality of the learning.

Keywords: language learning, speech perception, speech categorization, auditory categorization, segmentation

Supplemental materials: <http://dx.doi.org/10.1037/xhp0000067.supp>

A radio news show broadcast in an unfamiliar language seems to race by, giving the impression that the language uses very long words and that the broadcaster barely pauses for breath. This impression arises, in part, because the acoustic speech signal does

not consistently highlight linguistically significant units with pauses like the spaces that mark words in text (Cole & Jakimik, 1980). Through experience, listeners must discover a constellation of diagnostic acoustic and statistical cues such as prosody, stress patterns, allophonic variation, phonotactic regularities, and distributional properties that support word segmentation (see Jusczyk, 1999). Complicating matters for adults listening to speech in a non-native language, native-language segmentation cues influence adults' evaluation of non-native speech and may lead to inaccurate segmentation when the languages' cues do not align (e.g., Altenberg, 2005; Barcroft & Sommers, 2005; Cutler, 2000; Cutler, Mehler, Norris, & Segui, 1986; Cutler & Otake, 1994; Flege & Wang, 1990; Weber & Cutler, 2006).

Making sense of an unfamiliar, continuous acoustic stream like a foreign news broadcast is further complicated by the fact that the acoustics of neighboring speech sounds, syllables, and words do not stack neatly like adjacent pearls on a string (Hockett, 1955). Rather, they intermingle so that there is substantial acoustic variability in the realization of speech produced in different contexts (Fougeron & Keating, 1997; Moon & Lindblom, 1994). Additional acoustic variability arises from inherent differences across talkers (Johnson, Ladefoged, & Lindau, 1993; Peterson & Barney, 1952) and even from outside sources like room acoustics (Watkins, 2005; Watkins & Makin, 2007). As a result, the functional "units" of

Sung-Joo Lim, Department of Psychology, Carnegie Mellon University; Francisco Lacerda, Department of Linguistics, Stockholm University; Lori L. Holt, Department of Psychology, Carnegie Mellon University.

Sung-Joo Lim is now at the Max Planck Research Group "Auditory Cognition," Max Planck Institute for Human Cognitive and Brain Sciences.

This work was supported by awards to Francisco Lacerda and Lori L. Holt from the Riksbankens Jubileumsfond (K2003-0867); grants to Lori L. Holt from the National Science Foundation (BCS-0746067) and the National Institutes of Health (R01DC004674); and training grants to Sung-Joo Lim from the National Science Foundation (DGE0549352), the National Institutes of General Medical Sciences (T32GM081760), and the National Institute on Drug Abuse (5T90DA022761-07). The authors thank Frederic Dick, Jason Zevin, and Natasha Kirkham for stimulating discussions regarding the project.

Correspondence concerning this article should be addressed to Sung-Joo Lim, Max Planck Research Group "Auditory Cognition," Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1A, 04103, Leipzig, Germany. E-mail: sungjoo@cbs.mpg.de

language (such as phonetic categories or words) that must be discovered from the continuous sound stream vary considerably in their physical acoustic realization. To communicate effectively, listeners must come to treat these variable instances as functionally equivalent. To do so, they must discover linguistically relevant variability, generalize across linguistically insignificant variability, and relate this learning to new instances; they must learn to *categorize* speech (Holt & Lotto, 2010). The category learning that begins in infancy for the native language (e.g., Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Kuhl et al., 2006; Werker & Tees, 1983) may complicate listening to speech in a non-native language in adulthood because well-learned native categories may not align with categories of the non-native language (Best, 1995; Best, McRoberts, & Goodell, 2001; Flege, 1995). Such is the case in the classic example of native Japanese adults' difficulty with the English /r/ and /l/ (Goto, 1971; Iverson et al., 2003; Miyawaki et al., 1975).

But how do listeners discover the acoustic variability that is linguistically relevant while also discovering the cues that support segmenting what is linguistically relevant from continuous sound? These two learning challenges are inherently concurrent; in natural spoken language, listeners must discover functional units relevant to language from a fairly continuous spoken sound stream without a priori knowledge of the temporal window that characterizes the units (e.g., phoneme, syllable, word). Since the detailed acoustics of the units vary across instances as a function of context and other factors, learners must generalize beyond highly variable experienced acoustics to new instances and, ultimately, relate these units to referents in the environment. Klein (1986) refers to this as the adult language learner's "problem of analysis" (p. 59). We know very little about speech category learning in this richer context because laboratory studies typically investigate speech category learning across isolated, individuated sounds (e.g., syllables or words) that are not embedded in fluent, continuous sound (e.g., Grieser & Kuhl, 1989; Ingvalson, Holt, & McClelland, 2012; Kuhl et al., 1992; Lim & Holt, 2011; Lively, Logan, & Pisoni, 1993; Werker & Tees, 1983, 1984).

The present research addresses how adults contend with the problem of analysis by placing listeners in a toy model of the language-learning environment—an immersive video game in which novel, continuous sound embedded with functionally relevant, though acoustically variable, category instances serves to support adaptive behavior through its relationship with visual referents. In this way, we examine auditory category learning in the context of continuous sound.

It is important that the sounds experienced in the video game be as unfamiliar as possible in order to control and manipulate listeners' histories of experience. In Experiment 1, this was accomplished using a natural language (Korean) unfamiliar to listeners. In Experiments 2 and 3, we exerted even stronger control over listeners' familiarity with the sounds by creating a completely novel soundscape. To do so, we exacted an extreme acoustic manipulation, spectral rotation, on English sentences (Blessner, 1972). This rendered the speech wholly unintelligible while preserving the spectrotemporal acoustic complexities that characterize the multiple levels of regularity (and variability) present in natural speech. Specifically, we spectrally rotated each utterance so that the acoustic frequencies below 4 kHz were spectrally inverted. In contrast to natural speech (including the Korean speech in Exper-

iment 1), these spectrally rotated sounds had no acoustic energy above 4 kHz. Although spectral rotation preserves some of the acoustic regularities present in natural speech, listeners do not readily map rotated speech to existing language representations (Blessner, 1972). Using these highly unusual acoustic signals that nonetheless capture the spectrotemporal regularities and complexities of speech, we investigated the extent to which learning in the context of simultaneous category learning and segmentation challenges generalizes to nonspeech auditory signals that are impossible productions for a human vocal tract (see Scott, Blank, Rosen, & Wise, 2000).

In the present study, our aim is to determine whether listeners discover a particular functional unit—embedded target words—within continuous sound streams. We use naturally spoken sentences, each with one of four target words embedded. The sentences are recorded multiple times so that, across recordings, the acoustics are variable, with considerable coarticulation and natural variation in rate and amplitude. Critically, even the target words are acoustically variable across utterances. Experiment 1 listeners experience unfamiliar non-native Korean sentences recorded by a native Korean speaker. For Experiments 2 and 3, we begin with English sentences spoken by a native English talker. We then spectrally rotate these sentences, rendering them unintelligible. This approach to creating a novel soundscape of continuous sound embedded with to-be-learned auditory categories preserves the variability and regularity of natural spoken language across sounds that are novel and unintelligible to naïve listeners. Across all three experiments, the challenge for listeners is to discover the functional equivalence of the acoustically variable target words from the continuous stream of unfamiliar, non-native, and unintelligible nonspeech sounds without a priori knowledge of the temporal window that characterizes this unit; that is, they must discover the new categories from continuous sound.

One means by which learners may do so is via the relationship of the sound categories to visual referents in the environment. Visual referents co-occurring with sound regularities are known to support both speech categorization and segmentation (Thiessen, 2010; Yeung & Werker, 2009), but it is as yet unclear the extent to which they may support discovery of auditory categories from continuous sound streams because studies have typically paired a single presentation of a visual referent with the onset of a corresponding sound (sometimes with a slight temporal jitter) or systematically presented referents with an isolated word or syllable (Cunillera, Laine, Càmarà, & Rodríguez-Fornells, 2010; Thiessen, 2010; Yeung & Werker, 2009). Nonetheless, it may be hypothesized that the presence of co-occurring visual referents may support category learning in the context of continuous sound by signaling the distinctiveness of acoustically similar items across referents (Thiessen, 2010; Wade & Holt, 2005; Yeung & Werker, 2009) and/or the similarity of acoustically distinct items paired with the same referent.

Unlike previous studies that have temporally synchronized audiovisual presentation, in the present study, the appearance of a visual referent instead coincides with what can be thought of as a short "paragraph" of speech (or nonspeech, in the case of Experiments 2 and 3), with a target word appearing at different positions within the constituent sentences. To illustrate, imagine a speaker holding a book and saying, "Hey there, have you seen my *book*? It is a *book* I checked out from the library. It is the *book* with a red cover." The visual

referent, present throughout the speech, serves as a correlated visual signal for the acoustically variable instances of *book* peppering the continuous acoustic stream. The approach we take in the present experiments is similar. The present stimulus paradigm is distinct from previous research in its approach to auditory–visual correspondence and moves a step beyond studies of how learners track consistent audiovisual mapping across multiple encounters with established sound units (e.g., Smith & Yu, 2008; Yu & Smith, 2007) because the units of sound, their temporal extent, and their position within the continuous sound stream are entirely unknown to learners.

We present the continuous sounds and their visual referents to adult learners in the context of an immersive video game (Leech, Holt, Devlin, & Dick, 2009; Lim & Holt, 2011; Liu & Holt, 2011; Wade & Holt, 2005). This environment allows for strict experimental control while more closely modeling the natural learning environment’s converging multimodal information sources, the need to use this auditory information to guide action, and the internal feedback present from successfully making predictions about upcoming events. It is a considerable departure from category learning tasks that make use of overt perceptual decisions and/or explicit performance feedback (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Chandrasekaran, Yi, & Maddox, 2014; Lively et al., 1993; Logan, Lively, & Pisoni, 1991) and also from entirely passive exposure paradigms in which listeners hear instances or streams of sounds without any overt behavior directed at the stimuli (e.g., Maye & Gerken, 2000, 2001; Saffran, Aslin, & Newport, 1996). To succeed within the video game, listeners must use sound to predict events and navigate the environment. The extent to which their predictions about the sounds are met with success or failure in meeting the goals of the game provides a form of internal feedback that may be supportive of learning (Lim & Holt, 2011; see Lim, Fiez, & Holt, 2014, for a review). This approach provides a means of investigating incidental, active learning that may more closely model some aspects of learning in the natural environment (see Lim & Holt, 2011; Lim et al., 2014).

Previous research has demonstrated that, within this game, adults readily learn to categorize nonspeech sounds with a complex acoustic structure (Leech et al., 2009; Liu & Holt, 2011; Wade & Holt, 2005) that are not learned readily through passive exposure (Emberson, Liu, & Zevin, 2013; Wade & Holt, 2005). Moreover, adults improve in second-language speech categorization when non-native syllables are presented in the game environment (Lim & Holt, 2011). However, as is typical of category learning experiments, each of these prior studies has simplified the category learning challenge by presenting listeners with isolated, segmented category exemplars in training.

In marked contrast, the present study provides no individuation; the visual referents are presented along with continuous streams of sound. Embedded within the continuous sound stream, exemplars from one of four acoustically variable sound categories (the target words) are associated with each visual referent. The temporal duration of the targets is unknown to participants and variable, as is the position of the targets in the continuous sound stream. However, the stimulus inventory is constructed such that the target word, although acoustically variable, is the portion of the acoustic stream most reliably related to a particular visual referent. Though acoustically variable, in a relative sense the target words are islands of reliability within the highly variable continuous acoustic signal because they are the only segment of the continuous acoustic stream that co-occurs consistently with a particular visual referent.

We predict that listeners will be sensitive to input regularities at the largest possible grain size that gives them predictive control over their environment (Ahissar & Hochstein, 2002). In the case of the artificial sound inventories exploited in the present experiments, these are the acoustically variable, yet relatively more reliable, target words. Despite the lack of an overt categorization or segmentation task and corresponding explicit feedback, we hypothesize that experience with the acoustic regularities characterizing targets will lead listeners to learn to categorize familiar sentences possessing the target words, to generalize to unfamiliar sentences that include the target words, and to generalize this knowledge to isolated instances of the target words never heard previously, thus demonstrating a beginning ability to segment the relevant acoustics from continuous sound and generalize category learning to new instances.

Experiment 1

In this experiment, we examined whether native English adult listeners can learn to segment target words appearing within a natural but unfamiliar spoken language—Korean. English and Korean are quite distinct phonologically, morphologically, and grammatically. Korean is a nontonal language, but in contrast with English, it has a triple consonant system, including soft, hard nonaspirated, and hard aspirated consonant contrasts, as well as triple contrastive ranges of vowel sounds involving monophthongs and two different kinds of diphthongs. In addition, Korean is quite morphologically complex compared to English, and the word order is typically subject–object–verb compared to subject–verb–object in English (Grayson, 2006). There are large phonotactic differences in the two languages, particularly in interpretations of consonant clusters and markers of word boundaries (Kabak & Idsardi, 2007). The two languages also differ in rhythmic structure; Korean is often described as syllable timed, whereas English is a stress timed language. Thus, engaging English monolingual listeners with unfamiliar, and linguistically quite distinct, Korean provides a means of presenting an ecologically realistic instantiation of the challenge of learning speech categories from continuous sound in adult language learning.

Method

Participants. Thirty-two monolingual English participants from Carnegie Mellon University were recruited for \$30 compensation. All participants were unfamiliar with the Korean language; they had neither studied nor been exposed to spoken Korean. All reported normal hearing. An additional 11 participants with the same characteristics served as naïve listeners in a brief test of the stimulus materials.

Stimuli. There were four to-be-learned Korean target words translated as English *red* ($/p^*alkan/$), *blue* ($/p^haran/$), *green* ($/tʃ^h-orok/$), and *white* ($/hayan/$),¹ each uttered in six different sentences in a natural, coarticulated manner by a native Korean female speaker (Sung-Joo Lim) in a sound-isolated room (16 bit, 22.05 kHz). Target words were uttered in six different sentence

¹ Korean has a three-way voiceless stop consonants contrast, differing from English’s two-way contrast. *Red* spoken in Korean ($/p^*alkan/$) has a Korean fortis stop, whereas *blue* ($/p^haran/$) and *green* ($/tʃ^h-orok/$) have Korean aspirated stops in the word-initial position (see Cho, Jun, & Ladefoged, 2002, for phonetic notations).

contexts. Each target word–sentence pairing was uttered four times to further increase acoustic variability. Across this 96-sentence training stimulus set (6 sentences \times 4 words \times 4 utterances), the four target words defined the to-be-learned categories and, due to their acoustic variability within category, modeled the challenge of learning functional equivalence classes. Moreover, since the target words appeared in sentence-initial, -medial, and -final positions in fluent speech and were never presented to listeners in isolation, the stimuli modeled the challenge of learning categories from continuous sound. Three additional sentences with the target words embedded and an isolated utterance of each target word were recorded to test generalization at posttest. Table 1 lists the sentences. On average, sentence stimuli were 2.25 s [minimum = 1.69 s, maximum = 3.20 s] in duration, and isolated target words were 0.94 s [minimum = 0.80 s, maximum = 1.05 s] in duration.

Eleven native English participants were recruited separately and were asked to freely sort the stimuli into four groups in a consistent manner, without familiarization or instructions on how to base their decisions. This provided a baseline measure of naïve English listeners' categorization of the Korean sentences according to the target word. After sorting the sentences, participants were familiarized with the sentences (a total of 60 sentence exposure trials) through passive listening for 5 min and tested again. The naïve listeners exhibited above-chance (25%) consistency in their sorting, $M = 36.2\%$, $SD = 7.2\%$, $t(10) = 5.14$, $p \leq .001$, Cohen's $d = 1.55$, before the familiarization phase. However, there was no change in sorting consistency following additional familiarization, $M_{change} \leq 1\%$, $F(1, 10) = 0.345$, $p > .5$, $\eta_p^2 = 0.033$. This reveals that native English listeners were able to discover some regularities present in the continuous stream of Korean sentences through exposure but that brief passive listening did not additionally boost sorting performance. These data provide a baseline for comparison of learning within the video game paradigm.

Procedure. Training was accomplished using the Wade and Holt (2005) video game paradigm. In the game, subjects navigated through a pseudo-three-dimensional space, encountering four animated "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 starting position and assignment counterbalanced across subjects). 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 (see Figure 1) and was counterbalanced across participants.

Each alien was associated with one target word (randomized in the assignment across participants) such that each time it appeared, multiple randomly selected exemplars (from the set of 24 sentences with the target word embedded; see Table 1) were presented in a random order through the duration of the alien's appearance until the participant completed a capturing or destroying action. Thus, subjects heard continuous speech, with target words associated with the visual image of the alien, the spatial quadrant from which the alien originated, and the motor–tactile patterns involved in capturing or destroying the alien. This exposure was fairly incidental; the sounds were of no apparent consequence to performance, and participants were instructed only to navigate the game. Early in game play, aliens appeared near the center of the screen and approached slowly, with alien appearance synchronized to the

onset of the entire utterance of a randomly drawn sentence. This provided participants time to experience the rich and consistent regularities between the visual referents and the sound categories. There was a great deal of visual and spatial information with which to succeed in the task, independent of the sound categories. Participants were not informed of the nature of the sounds nor their significance in the game. Other sound effects (including continuous, synthetic background music) were also present.

As the game progressed, the speed and difficulty of the required tasks increased so that quick identification of approaching aliens by means of their characteristic sounds was, while never required or explicitly encouraged, of gradually increasing benefit to the player. As the game difficulty increased, continued progress was only possible with reliance upon auditory cues. At higher levels of the game, players could hear the aliens before they could see them and, thus, if they had learned about the relationship of the acoustically variable sounds to the visuospatial characteristics of the alien, they could use the acoustic information to orient behavior more quickly to succeed in the goals of the game. At the highest game levels, targeting became nearly impossible without rapid sound categorization, which predicted the alien and its quadrant of origin and provided participants a head start on navigating and orienting action in the right direction. In this way, sound served as a cue to predict appropriate action, thus encouraging sound category learning through its utility for functioning in the environment, without requiring overt categorization responses. Of special note with regard to our research aims, each appearance of an alien triggered continuous presentation of sound(s) from the category. There was no indication of the relevant functional units (or that there were relevant units to be discovered) or the temporal window across which they unfold and no explicit feedback about sound segmentation or categorization.

Participants played the video game for two 50-min sessions separated by a 10-min break. An explicit posttest assessed learning and generalization immediately after training. Participants responded to 10 stimuli for each target word (four repetitions each): one randomly chosen utterance of each of the six familiar sentences experienced in training and the four novel test stimuli given in Table 1. Participants saw a game screen with the four aliens, each positioned in its typical quadrant. On each trial, a single randomly drawn stimulus was played repeatedly for as long as 5.5 s. Participants used the arrow keys as a means of classifying the given sound stimulus. It is of note that while most categorization learning studies use relatively comparable training and testing experiences (e.g., highly similar explicit categorization tasks with the same category mapping labels or response keys), the posttest in the current study is highly distinct from the incidental nature of sound categorization in the video game training experience. Therefore, we analyzed listeners' response patterns across consistent response–sound stimulus mappings to measure the overall correct categorization performance.

Both the video game training and explicit posttest categorization tasks were presented on the center of a computer monitor (600 \times 600 pixels) mounted on the wall of a sound-attenuated booth. Participants used a keyboard to interact with the game. All sounds were presented through headphones at a comfortable listening level (approximately 70 dB). The mapping between the alien creatures and the target word categories was randomized across participants, thus destroying the color match between the visual

Table 1

Korean Sentences Used in Video Game Training and Reserved as Novel Generalization Stimuli at Posttest

Training		Test	
1. 총으로 [] 표적을 쏘아라.	Shoot the [] target with a gun.	1. []	[]
2. 적은 [] 색이다.	The color of the enemy is [].	2. 어느것이 [] 물체냐?	Which object is []?
3. [] 대상에 유의하라.	Look out for the [] object.	3. 지금 []색을 골라라.	Choose color [] now.
4. [] 외계인을 보아라.	Watch for [] aliens.	4. [] 목표물이다.	It is a [] target.
5. 나쁜것은 [] 물체다.	The bad one is a [] thing.		
6. 지금 오는것은 [] 침입자이다.	[] invaders are coming now.		

Note. The brackets denote the placement of Korean target words, translated as blue, white, green, and red. There were four unique recordings of each sentence to increase acoustic variability.

appearance of the alien and the target word meaning (in Korean) across participants.

Data from additional naïve native English participants served as a baseline accuracy measure in categorizing the Korean sentences without training. Comparing baseline performance across, rather than within, participants ensured that the trained participants entered into training without an explicit indication of the significance of the sounds to the game, or the learning questions under investigation.

Results and Discussion

Despite the presence of acoustic variability introduced by the stimulus materials, the lack of a consistent temporal window across which to predict the target word's appearance in a sentence, and the absence of performance feedback or an explicit categorization or segmentation task, participants categorized Korean target words above chance (25%) at posttest. One sample t test revealed that Experiment 1 listeners' posttest categorization performance was significantly different from chance [$M = 53.5\%$, $SD = 27.0\%$, $t(31) = 5.98$, $p \leq .001$, $d = 1.06$].² Moreover, Experiment 1

listeners were significantly more accurate than naïve participants in sorting Korean stimuli without training. An independent samples t test on the categorization performance of Experiment 1 and naïve participants' baseline sorting performance revealed a significant group difference in categorizing Korean stimuli, $M_{diff} = 17.2\%$, $t(41) = 2.09$, $p = .043$, $d = 0.73$. We further investigated whether learning was observed for familiar and novel generalization sentences as well as isolated instances. One sample t tests against chance (25%) revealed that Experiment 1 participants reliably categorized both familiar training sentences, which participants experienced during the video game training [$M = 52.4\%$, $SD = 26.9\%$, $t(31) = 5.76$, $p \leq .001$, $d = 1.02$], as well as novel generalization sentences [$M = 53.9\%$, $SD = 28.1\%$, $t(31) = 5.84$, $p \leq .001$, $d = 1.03$] and isolated Korean target words [$M = 59.6\%$, $SD = 29.6\%$, $t(31) = 6.61$, $p \leq .001$, $d = 1.17$] that participants never experienced during training. This indicates that participants were able to generalize what they learned about the acoustically variable functional units—the target words—from continuous speech (see Figure 2).

We further examined whether listeners' categorization performance differed across the placement of the target words appearing in sentences as well as isolated instances. A repeated-measures analysis of variance (ANOVA) on participants' posttest categorization performance of sentence-initial, -medial, and -final, as well as isolated, target words revealed a significant effect of target word placement [$F(3, 93) = 6.74$, $p \leq .001$, $\eta_p^2 = 0.179$]. Post hoc comparisons revealed that categorization performance for sentence-initial target words was equivalent to that for isolated target words [$M_{diff} = 1.3\%$, $t(31) = 0.66$, $p > .5$, $d = 0.12$] and significantly more accurate than categorization of target words in the sentence-medial [$M_{initial-medial} = 7.5\%$, $t(31) = 2.58$, $p = .015$, $d = 0.46$] or -final positions [$M_{initial-final} = 8.8\%$, $t(31) = 3.16$, $p = .003$, $d = 0.56$], which did not differ from one another [$M_{medial-final} = 1.3\%$, $t(31) = 0.83$, $p > .4$, $d = 0.15$]. One factor that may have influenced this pattern of results is the degree of acoustic variability across utterances; sentence-initial and isolated target words may have been uttered with somewhat less coarticu-

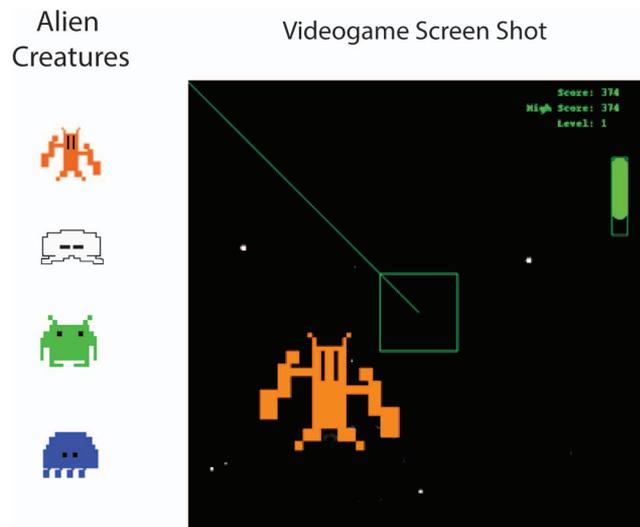


Figure 1. Screenshot of the video game training. As one of four alien creatures (shown in the left panel) approaches in each trial, listeners need to make correct motor actions associated with the alien. See the online article for the color version of this figure.

² Sixteen of the participants heard Korean speech spoken in a more exaggerated manner in order to model infant-directed speech input. The remainder of the participants heard the same Korean sentences spoken by the same talker in adult-directed speech, without exaggeration. The exaggeration had no impact on learning [$M_{diff} = 3.79\%$, $t(30) = 0.39$, $p > .5$] or latency of posttest responses [$M_{diff} = 241$ ms, $t(30) = 1.40$, $p = .17$], so all results are collapsed across groups.

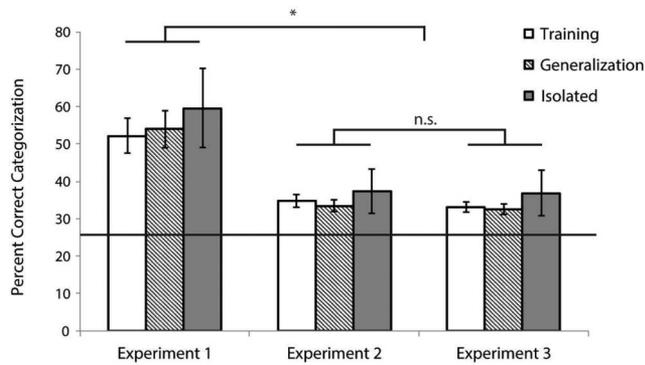


Figure 2. Average percent correct categorization for familiar training, novel generalization, and novel isolated target word stimuli in Experiment 1 (fluent Korean speech) and Experiments 2 and 3 (spectrally rotated English speech) (* indicates that $p \leq .001$). The line at 25% indicates chance-level performance. The learning exhibited across all experiments and all test stimuli was significantly above chance ($p \leq .001$). Error bars indicate standard errors of the mean.

lation than sentence-medial or -final targets, especially since it is impossible to have grammatically valid Korean sentences in which the target word appears at the end of the sentence (see Training 2 and 5 and Test 2 from Table 1). This may have exaggerated sentence-final target coarticulation. It is also possible that recognition may have been facilitated by the brief, natural pause that would precede target words in the sentence-initial position. Although the repeated presentation of sentences reduced demands on short-term auditory memory, a general primacy bias may have facilitated categorization of target words located at utterance boundaries (Aslin, Woodward, LaMendola, & Bever, 1996).

It is of note that performance varied across target words. A repeated-measures ANOVA revealed a main effect of the target word, $F(3, 93) = 2.986$, $p \leq .035$, $\eta_p^2 = 0.088$, with about a 5% disadvantage in categorizing Korean *blue* ($p^h\text{aran}$) and *red* (p^*alkan) relative to the other targets. This may be due to the relatively greater phonetic acoustic similarity of these targets.¹ Alternatively, it may indicate that listeners relied solely on the initial consonant rather than learning to segment the two-syllable word unit from speech. However, arguing quite strongly against this interpretation, there was no difference in listeners' categorization performance across isolated target words [$F(3, 93) = 1.152$, $p = .332$, $\eta_p^2 = 0.036$].

In sum, incidental learning within the video game paradigm was sufficient to induce reliable categorization of acoustically variable functional units from continuous sound, even without explicit categorization or feedback, and without exact temporal synchrony of auditory category exemplars and visual referents. Based on listeners' ability to generalize learning to novel sentences with target words embedded and also to isolated instances of the target word, we conclude that listeners began to discover the temporal grain size in the continuous acoustic input that granted them predictive control over the environment—the target words. Though acoustically variable, the target words were the window of acoustic information within the continuous sound that best correlated with the appearance of specific visual referents.

Experiments 2 and 3

Although listeners in Experiment 1 were unfamiliar with Korean, it is nonetheless possible that they were able to exploit commonalities between Korean and English (e.g., overlapping phonetic information) for category learning. This possibility is supported by the above-chance baseline ability of naïve listeners to sort the Korean sentences. Thus, in Experiments 2 and 3, we eliminated the potentially buttressing effects of language similarities by training adults with English speech signals radically manipulated using spectral rotation (Blessner, 1972). This signal processing technique preserves much of the acoustic regularity and variability of speech but renders the sentences wholly unintelligible. Since these signals are not perceived as speech, another aim of the experiments was to examine whether the learning observed in Experiment 1 generalizes to nonspeech acoustic signals.

Method

Participants. Forty and 37 native Swedish adults from Stockholm University participated in Experiments 2 and 3, respectively. Participants volunteered in return for a light meal and two cinema tickets. All reported normal hearing.

Stimuli. Training and generalization sentences similar to those of Experiment 1 were used, preserving the variation of the target word position within the sentences (sentence-initial, -medial, and -final positions). There were four to-be-learned target words (*red*, *green*, *blue*, and *white*), each spoken in English in a natural, coarticulated manner and uttered in six different digitally recorded sentences (16 bit, 22.05 kHz) by a monolingual English female talker (Lori L. Holt) in a sound-isolated room (see Table 2). The average duration of sentence stimuli was 1.31 s [minimum = 0.93 s, maximum = 1.78 s], and that of the isolated target words was 0.54 s [minimum = 0.46 s, maximum = 0.64 s]. Measures of the acoustic variability present across the multiple stimulus contexts are presented in the Appendix.

Beginning from these natural utterances, we exacted an extreme acoustic manipulation to render the speech unintelligible while preserving the spectrotemporal acoustic complexities that characterize the spoken language-learning environment. Each utterance was spectrally rotated using a digital version of the technique described by Blessner (1972). The sounds were low pass filtered to remove acoustic energy above 4 kHz, and the remaining frequen-

Table 2

Sentences That Served as the Basis for Spectrally Rotated Training and Novel Generalization Test Stimuli

Training	Test
1. Shoot the [] one.	1. []
2. [] is an enemy.	2. Which guy is []?
3. Look out for the [].	3. Choose [] now.
4. Watch for [] aliens.	4. [] targets now.
5. The bad guy is [].	
6. [] invaders are coming.	

Note. The brackets denote the placement of keywords (blue, white, green, and red).

cies below 4 kHz were inverted in the spectrum.³ The result is much like an unintelligible “alien” language that possesses the temporal and spectral complexity of ordinary speech. Blesser (1972) reported that it took intensive training, over a period of weeks, for listeners to extract meaning from rotated speech. Thus, listeners cannot readily map the acoustic patterns of spectrally rotated speech to existing language representations.

To test this explicitly with our own materials, eight naïve native English participants with normal hearing attempted to categorize the rotated speech target words used in training and testing. Target words were presented in isolation, segmented from the sentences.⁴ Even with instructions that the sounds originated from the English target words *red*, *green*, *blue*, and *white*, participants’ mean accuracy in this four-choice categorization test was no different from chance [$M = 27.0\%$, $SD = 8.4\%$, $t(7) = 0.659$, $p \geq .5$, $d = 0.23$]. Thus, spectral rotation sufficiently distorts the acoustic signal to prevent listeners from hearing it as speech or from using it to access speech categories or words, even when the target words are extracted from the continuous sound stream and presented in isolation. This is consistent with previous research reporting that passive listening to rotated speech does not engage left anterior superior temporal sulcus (STS) regions thought to be drawn online by intelligible speech (Scott et al., 2000). To further limit the possibility that the spectrally rotated English speech was mapped to existing English language representations, we conducted the experiment in Sweden with native Swedish-speaking participants. During the experiment, the experimenter never indicated the sounds’ origin as English or as speech. Figure 3 shows representative stimulus spectrograms and waveforms of an original utterance and its spectrally rotated counterpart. The sounds are available online as supplemental material.

Procedure. The video game procedure was identical to that of Experiment 1 with a slight variation in stimulus presentation. In Experiment 2, a randomly selected utterance of a single sentence possessing the associated target word was presented repeatedly on

each appearance of an alien. In Experiment 3, multiple, randomly selected utterances of different sentences were presented in a random order on each alien appearance, as in Experiment 1. Thus, the input in Experiment 2 modeled repeated instances of the exact same sentence (e.g., *Shoot the blue one! Shoot the blue one! Shoot the blue one!*), with target words in the same position and possessing identical sentence acoustics within an event, i.e., the appearance of a particular alien. Across events, however, the sentences varied and the acoustics of the target words were highly variable. Experiment 3 input was characterized by acoustic variability of target words and variability in target word position within an event, as well as acoustic variability across events (e.g., *Shoot the blue one! Blue invaders are coming! Look out for the blue!*).

In all, these materials presented a challenging learning environment. As in Experiment 1, there was considerable acoustic variability across target words, and they occurred embedded in continuous sound streams at unpredictable positions. The temporal window (word) defining the target was unknown to participants, who had to discover the predictive temporal window from continuous acoustics. Adding to the learning challenge in Experiments 2 and 3, there was no overlap with English language knowledge (potentially available in Experiment 1’s Korean materials); the lower-frequency acoustics informative to the identity of the target words were remapped in higher-frequency bands from the spectral rotation manipulation. Thus, the reversed mapping of acoustic signals of spectrally rotated speech significantly deviated from the acoustic regularities of speech or natural sounds more generally.

Participants were instructed briefly (in Swedish) on how to play the video game, but no mention was made about the significance of the sounds to the game or their relationship to English, speech, or the learning questions under investigation.

Results and Discussion

As in Experiment 1, listeners performed significantly above chance (25%) in the posttest categorization of the rotated speech after just a total of 100 min of game playing within a single experimental session. Experiment 2 participants, for whom a single spectrally rotated sentence repeated within an event, performed significantly above chance on posttest categorization of the familiar sentences [$M = 34.8\%$, $SD = 11.0\%$, $t(39) = 5.67$, $p \leq .001$, $d = 1.26$] and generalized to novel never-before-heard spectrally rotated sentences possessing the same target word [$M = 33.5\%$, $SD = 9.9\%$, $t(39) = 5.42$, $p \leq .001$, $d = 0.86$]. Perhaps most striking, performance on *isolated* target words never heard in training was significantly above chance [$M = 37.4\%$, $SD = 18.4\%$, $t(39) = 4.26$, $p \leq .001$, $d = 0.93$] (see Figure 2). Above-chance categorization of the isolated target words demonstrates that listeners were able to extract the equivalence class categories associated with target words from the continuous sound. Moreover, since the instances of isolated targets were acoustically

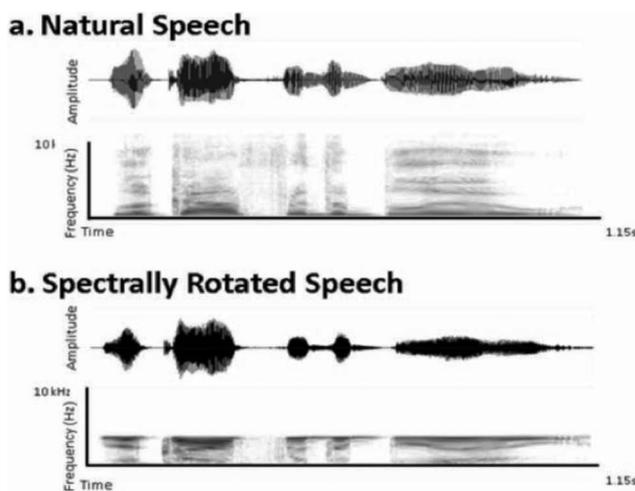


Figure 3. Spectrogram (Time \times Frequency) and waveform (Time \times Amplitude) illustrations of a single utterance of “Look out for the blue.” The upper panel illustrates the original recording of the utterance. The bottom panel shows the spectrally rotated version of the same utterance used in training.

³ The original frequencies between 0 and 4,000 Hz (f_i) were remapped linearly according to $f_{i_rotated} = 4,000 \text{ Hz} - f_i$.

⁴ Note that since we only tested performance on the isolated target words, the pilot listeners may have been disadvantaged by the lack of potentially supportive information from context, which is known to affect the intelligibility of fluent speech (e.g., Pickett & Pollack, 1963; Pollack & Pickett, 1964).

different from those experienced in training, their above-chance performance indicates that listeners could relate the acoustically variable segments within the continuous sentences presented during learning to the acquired equivalence classes—a step toward successful segmentation of category exemplars.

Regardless of whether the target word appeared in the initial, medial, or final position in the sentence, categorization was above chance [initial: $M = 35.0\%$, $SD = 11.2\%$, $t(39) = 5.62$, $p \leq .001$, $d = 0.89$; medial: $M = 30.5\%$, $SD = 9.4\%$, $t(39) = 3.68$, $p \leq .001$, $d = 0.58$; final: $M = 36.2\%$, $SD = 12.5\%$, $t(39) = 5.68$, $p \leq .001$, $d = 0.90$]. However, a repeated-measures ANOVA revealed a significant main effect of target word position in the sentence (initial, medial, and final), $F(2, 78) = 6.17$, $p = .003$, $\eta_p^2 = 0.14$]; categorization performance of target words appearing in either sentence-initial or -final positions was equivalent to categorization of the isolated target words [$F(2, 78) = 0.64$, $p > .5$, $\eta_p^2 = 0.016$], all of which were significantly better than words in the sentence-medial position [$F(3, 117) = 4.52$, $p = .005$, $\eta_p^2 = 0.104$]. Thus, there was a slight disadvantage to learning targets in the medial position, perhaps due to greater acoustic variability drawn from coarticulation of sentence-medial productions or to a disadvantage due to the relatively privileged segmentation of the target words appearing at the edges of an utterance, as speculated for Experiment 1.

Similarly, Experiment 3 listeners' overall posttraining categorization performance was significantly greater than chance [$M = 32.9\%$, $SD = 7.6\%$, $t(36) = 6.31$, $p \leq .001$, $d = 1.04$]. This was true for familiar stimuli heard in training [$M = 33.1\%$, $SD = 8.1\%$, $t(36) = 6.09$, $p \leq .001$, $d = 1.00$], novel sentences [$M = 32.5\%$, $SD = 8.2\%$, $t(36) = 5.60$, $p \leq .001$, $d = 0.92$], and novel isolated target words [$M = 36.8\%$, $SD = 15.6\%$, $t(36) = 4.60$, $p \leq .001$, $d = 0.76$] (see Figure 2). Experiment 3 participants' performance was statistically indistinguishable from that of Experiment 2 listeners, $t(75) = 0.69$, $p = 0.49$, $d < 0.20$, indicating that the additional acoustic variability present in Experiment 3 neither helped nor hurt learning.⁵

Experiment 3 listeners categorized the target words placed in the sentence-final position as well as they did those presented in isolation [$t(36) = 1.37$, $p \geq .18$] and significantly more accurately than sentence-medial words [$F(2, 72) = 5.23$, $p = .008$, $\eta_p^2 = 0.127$]. However, although the sentence-initial word categorization was indistinguishable from sentence-final or isolated word categorization [$F(2, 72) = 1.70$, $p = .190$, $\eta_p^2 = 0.045$], it was not significantly different from sentence-medial word categorization [$t(36) = 1.74$, $p = .091$, $d = 0.29$]. It is possible that the additional within-trial acoustic variability may have blurred the boundaries of each sentence within a trial, mitigating the word-initial advantage observed in the other experiments.

The chance-level matching performance of naive participants strongly suggests that spectral rotation distorted the acoustics of English speech sufficiently to prevent lexical access. In keeping with this, posttest categorization performance was similar across participants regardless of the counterbalanced mapping between the target words and visual aliens. That is, categorization performance of participants who experienced target words that mismatched the alien colors (e.g., a *red* alien paired with spectrally rotated sentences embedded with the target word *blue*) did not differ from listeners who experienced target words that matched the alien colors [$M_{diff} = 0.9\%$, $t(75) = 0.45$, $p > .5$]. There were

also no differences in performance across target words [Experiment 2: $F(3, 117) = 1.101$, $p = .352$, $\eta_p^2 = 0.027$; Experiment 3: $F(3, 108) = 0.690$, $p = .560$, $\eta_p^2 = 0.019$].

However, in comparison to Experiment 1 listeners' learning of unfamiliar Korean speech, learning of spectrally rotated English materials in Experiments 2 and 3 was significantly worse [$M_{Korean} = 53.5\%$; $M_{Rotated} = 33.6\%$; $t(107) = 5.82$, $p \leq .001$, $d = 1.07$; see Figure 2]. Also, listeners who heard spectrally rotated speech responded more slowly to posttest categorization items than those learning Korean targets [$M_{Korean} = 2,003$ ms; $M_{Rotated} = 2,420$ ms; $t(107) = 2.97$, $p = .004$, $d = 0.61$]. This pattern of results may indicate a greater advantage for learning speech materials compared to nonspeech sounds created by spectrally rotating speech. Issues regarding the domain generality of learning are discussed in more detail below.

General Discussion

Perceptual systems must discover behaviorally relevant regularities associated with objects and events embedded in highly variable, and continuous, sensory input. Spoken language is an example. Natural speech possesses considerable acoustic variability such that no unique acoustic signature distinctly defines linguistically relevant units like individual phonemes, syllables, or words. Further complicating speech category learning, acoustic information for these units must be drawn from a nearly continuous acoustic stream without strong physical markers, like silence, to signal unit boundaries in time. There is a rich literature addressing speech category learning across development and into adulthood, and there is a growing body of research on auditory category learning of nonspeech signals (see Holt, 2011, for a review). However, how listeners learn categories from imperfect sources of acoustic information simultaneously available at multiple temporal granularities in continuous sound input is an important, but unresolved, issue.

The present results demonstrate that listeners can recognize acoustically variable word-length segments embedded in continuous sound without prior knowledge of the functional units, their temporal grain size, or their position in the continuous input. Moreover, they can learn these categories in the context of a largely incidental task in which visual referents support learning. In the present study, the target words possessed great variability across spectral and temporal acoustic dimensions (see the Appendix), their position within sentences was unpredictable, and there were no a priori cues that the temporal grain size of “two-syllable” (Experiment 1) or “one-syllable” (Experiments 2 and 3) words was

⁵ Although there was no benefit of Experiment 3's additional variability on categorization, Experiment 3 listeners did make their categorization responses faster than Experiment 2 participants [$M_{diff} = 335.9$ ms, $t(75) = 2.08$, $p = .041$]. This trend was similar for both correct and incorrect trials [correct: $M_{diff} = 284.9$ ms, $t(75) = 1.83$, $p = .071$; incorrect: $M_{diff} = 374.9$ ms, $t(75) = 2.23$, $p = .029$]. Considering the modest level of categorization performance in the two experiments (about 34%), the small number of correct trials reduced the power, which may have resulted in a marginally significant response time difference for correct trials for Experiments 2 and 3. Owing to the between-subjects posttest design, it is not possible to conclude whether listeners in Experiment 3 were generally faster in providing responses or whether Experiment 3 training provided a greater benefit for faster recognition and segmentation of target words than Experiment 2.

significant (regularity at the whole-sentence, phrase, syllable, or phoneme level might have defined the sound category inventory). Our results indicate that participants capitalized on the temporal granularity that best supported behavior within the video game (Ahissar & Hochstein, 2002); only the target words reliably covaried with the visual referents and thus had predictive power in guiding relevant behaviors in the game.

Under these conditions, we observed category learning for non-native speech and even for unintelligible nonspeech stimuli created by warping English speech with spectral rotation. Spectral rotation eliminated similarities between English and Korean that may have supported learning non-native speech in Experiment 1 and created a soundscape in which a priori presumptions of the critical acoustic features or of the temporal granularity that best relates to language were minimized. Yet, learning was observed. This hints that domain-general processes may be involved.

Domain-General Learning?

Regardless of whether to-be-learned materials were speech or nonspeech, reliable learning was observed. Nevertheless, it is evident that there was a learning advantage for the natural Korean speech materials compared to the nonspeech signals created by spectrally rotating speech. However, it should be noted that because these comparisons were between groups, a priori between-groups differences unrelated to learning cannot be ruled out. The different extent of learning for speech and nonspeech might be interpreted to be an advantage for learning speech, as compared to nonspeech signals, with theoretical implications for the hypothesized advantage for learning relationships between speech and visual referents (e.g., Ferry, Hespos, & Waxman, 2010; Fulkerson & Waxman, 2007) and for a bias toward speech versus nonspeech in general (e.g., Vouloumanos & Werker, 2007).

However, in interpreting this speech versus nonspeech difference, it is important to note that native English listeners naïve to Korean were significantly above chance in sorting sentences with acoustically variable Korean target words, but naïve listeners were at chance in sorting spectrally rotated target words (even when informed about the target words' identity in unrotated English); thus, the baseline sorting ability for Korean was significantly higher than that of spectrally rotated speech, $M_{diff} = 9.3%$, $t(17) = 2.58$, $p \leq .020$, $d = 1.05$. Although the pilot tests of the stimuli were not conducted such that they invite a formal statistical comparison with participants who trained within the video game, it is informative to note that the baseline performance of naïve listeners sorting Korean stimuli ($M = 36.2%$ correct) was on par with the ultimate achievement of listeners who trained with spectrally rotated speech. Participants began with an advantage in discovering the Korean categories in Experiment 1 relative to listeners who learned spectrally rotated speech categories in Experiments 2 and 3. But what was the nature of this advantage?

One critical factor may relate to the acoustic differences between speech and its spectrally rotated complement. Spectral rotation preserves the natural spectrotemporal complexity of speech but inverts its frequencies. This results in a transformation of the highly informative lower frequencies (where most phonetic information is carried across languages) to higher-frequency bands. Although the frequency range of sounds spectrally rotated at 4 kHz remains well within the "sweet spot" (1,000–3,000 Hz) where the

human audiogram has its lowest thresholds (Fant, 1949), the relationship between the intensity of acoustic energy across frequencies in natural speech is reversed in rotated speech. Therefore, rotated speech violates natural sound signal statistics, thereby creating signals that are quite distinct from those the auditory system evolved to decode and that violate signal statistics consistent with listeners' long-term auditory experience. As such, it is reasonable to hypothesize that spectrally rotated speech may present greater learning challenges simply due to its highly unnatural acoustics and their potential impact on stimulus discriminability. Cunillera, Laine, et al. (2010), for example, report an effect of visual discriminability on cross-modally supported learning.

If the acoustic characteristics of spectrally rotated speech handicap learning due to greater demands on perceptual processing and not due to inherent learning differences for speech and nonspeech, then additional experience may push nonspeech category learning toward speech-like levels. We conducted a small follow-up study to address this possibility. Our aim was to determine whether the modest target word categorization accuracy observed for the nonspeech spectrally rotated signals represents an upper limit on nonspeech category learning. Six native English adult listeners played the video game with Experiment 3 input across six daily 1-hr sessions. Following this extended training, participants' categorization was significantly more accurate than that observed in Experiment 3 [$M_{6-hr} = 51.6%$, $SD = 20.5%$, $t(5) = 3.17$, $p \leq .025$, $d = 1.30$] and, impressively, one participant achieved 83.3% accuracy. These data indicate that greater experience can lead to highly accurate categorization of spectrally rotated speech despite the complex learning challenges presented by the signals. High levels of category learning can be achieved even with complex, continuous sounds that systematically correlate with referents in the environment (Colunga & Smith, 2002), even if they are not speech-like.

Another critical factor in considering the learning advantage we observed for speech versus nonspeech is listeners' history of experience. Although Korean was unfamiliar to our native English listeners, it shares common signal statistics identifying it as spoken language, and English and Korean overlap in some aspects of their phonology and phonotactics. It is easy to imagine how such sources of information may serve to guide and constrain listeners' hypotheses about the relevant functional units residing in continuous sound. Even recognition of the Korean sentences as speech in the broadest sense may support valid inferences about the temporal windows across which meaningful spoken language events take place (Poepfel, 2003). This may constrain the search for reliable windows of acoustic information in continuous sound. In contrast, in Experiments 2 and 3, listeners encountered the unusual signal statistics of spectrally rotated speech for the first time within the task. Thus, our speech and nonspeech stimuli also varied in the degree to which listeners had expertise relevant to parsing these sound streams. For this reason, it is necessary to be cautious in interpreting the present speech versus nonspeech learning differences as fundamental differences in learning because expertise with sound classes can modulate differences in speech and nonspeech processing. Leech et al. (2009), for example, found that a region of the left STS commonly considered to be speech selective (e.g., Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Binder et al., 2000; Scott et al., 2000) is recruited during passive listening to complex nonspeech sounds, but only among listeners who had

learned to categorize the nonspeech sounds. Although unintelligible, spectrally rotated speech is known to engage different patterns of cortical activation than speech signals among naïve listeners (Scott et al., 2000), the present results and those of Leech et al. (2009) suggest the possibility that spectrally rotated speech sounds might engage more speech-like cortical networks as listeners develop expertise in categorizing these signals. Speech and nonspeech signals can differ in more than their status as human-produced sounds or in their acoustics; consideration of listeners' expertise with the sounds is also a significant factor in interpreting domain generality versus specificity (Leech et al., 2009; see also Gauthier & Tarr, 1997).

Sensitivity to Input Regularities and Variability

Whether speech or nonspeech, we found that category learning generalized to novel utterances and to instances presented in isolation. Under the current approach, it is not possible to determine the specific acoustic dimensions that listeners relied on for categorization because, by design, acoustic dimensions freely varied to examine whether categories can be learned in the context of the natural acoustic variability present in fluent speech. Nevertheless, the results of Experiment 1 hint that learners were sensitive to acoustic similarity across target words; there was about a 5% disadvantage in categorizing Korean *blue* (*lp^haranl*) and *red* (*lp^{*}alkanl*) compared to the other targets. This disadvantage might be attributable to greater acoustic-phonetic similarity between these two words relative to the other targets in the inventory. Likewise, the consistent pattern of advantage for segmenting words in the sentence-initial and -final over -medial position across the experiments invites explanations related to regularities in acoustics, all of which favor the use of utterance edges for locating target words (Seidl & Johnson, 2006). Brief pauses at utterance boundaries played a critical role in facilitating categorization of words that aligned with utterance edges. Also, increased coarticulation may have created greater acoustic variability in the target words at the sentence-medial position relative to the -initial and -final positions. These possibilities suggest the crucial role of acoustic similarity and variability in contributing to category learning.

Another level of regularity worth mentioning is the manipulation of sentence structure in Experiments 2 and 3. Variability in speech input is beneficial for increasing sensitivity to phonemic distinctions in infants (Thiessen, 2007) and appears to promote learning and generalization of non-native speech categories in adults (e.g., Bradlow et al., 1997; Lively et al., 1993; Logan et al., 1991), at least in tasks with individuated phoneme classes. Experiments 2 and 3 used the same spectrally rotated speech stimulus inventory but varied in the presentation of sentences within an event (repeated sentence vs. multiple sentences). From a purely statistical stance, it might be argued that the additional variability introduced in the non-target-word acoustics of Experiment 3 would serve to highlight the relative reliability of the target word acoustics within an event, thus promoting discovery of the target word. However, this was not the case; learning was equivalent in Experiments 2 and 3. In understanding this, it may be useful to consider the task from the perspective of an optimal Bayesian model that would evaluate various possible input structures as candidate acoustic segments relevant to predicting the visual ref-

erent (Goldwater, Griffiths, & Johnson, 2009). When sentences repeated (Experiment 2), there was less information to process within an event. There were fewer possible sequential chunks to be evaluated, perhaps advantaging learning. However, in order to evaluate evidence for hypotheses, it was necessary to encounter multiple events. When multiple sentences occurred within each event (Experiment 3), there were perceptual processing demands in evaluating the possible sequential chunks, but it would have been possible to evaluate hypotheses even within an event. As such, it is possible that competing task demands masked effects of variability in the present approach. Future work might explicitly manipulate the information present within versus across appearances of a visual referent to determine these factors' independent contributions.

Learning Mechanisms and Implications

The category learning we observed occurred within the context of a challenging, immersive video game. Participants were not informed about the existence of auditory categories or even of the significance of sound to success within the game. However, the structure of the game was such that the to-be-learned categories embedded in continuous sound were the best predictors of specific aliens and the appropriate action. Thus, learning to treat sound category members as functionally equivalent served to support effective predictions about upcoming actions within the game. In this way, the task was not entirely passive or unsupervised. Feedback arrived in the form of success or failure in achieving goals, and there were multiple, correlated multimodal events and objects that covaried with category membership.

The success or failure of behaviors stemming from such predictions may engage intrinsically generated learning signals to a greater extent than passive, unsupervised training (see Lim & Holt, 2011; Lim et al., 2014). Passive exposure learning paradigms, often exploited in testing statistical learning of speech and other sounds among adults and infants (see Kuhl, 2004, for a review) may be limited in the extent to which they scale to learning challenges that incorporate more natural variability (Pierrehumbert, 2003). For example, with just passive exposure, infants fail to segment words of variable length from fluent speech streams (Johnson & Tyler, 2010; but see Thiessen, Hill, & Saffran, 2005, and Pelucchi, Hay, & Saffran, 2009, for evidence of the benefits of additional speech cues in aiding learning), and adults fail to learn functional equivalence classes for spectrally complex novel sounds that are learned readily in the present video game paradigm (Emberson et al., 2013; Wade & Holt, 2005).

In a recent study that also investigates how listeners learn from multiple sources of regularity simultaneously available at different temporal granularities in auditory signals, listeners were passively exposed to a continuous stream of unfamiliar, acoustically variable nonspeech "word"-level units comprising two very discriminable categories and two additional categories that highly overlapped in perceptual similarity (Emberson et al., 2013). In this way, as in the current study, listeners were confronted with simultaneous segmentation and categorization learning challenges. Although listeners were able to use the perceptually discriminable categories to discover the word-like units, passive listening to a continuous stream of these sounds for 7 min did not lead participants to discover the two experimenter-defined categories composed of

perceptually similar stimuli. These results suggest that there may be limitations on the power of passive exposure to drive learning at multiple levels of learning simultaneously. This is particularly interesting because these very same nonspeech stimulus categories were readily learned with the video game paradigm of the present study (Leech et al., 2009; Liu & Holt, 2011; Wade & Holt, 2005). The intrinsic reward of success in the game, of accurately predicting and acting upon upcoming events, and the rich multimodal correlations among actions, objects, and events present in the video game but absent in passive paradigms may be powerful signals to drive learning (Lim & Holt, 2011; Lim et al., 2014; Wade & Holt, 2005). In line with this possibility, a different type of incidental task (Seitz & Watanabe, 2009), whereby subthreshold task-irrelevant (thus, unattended) sounds are presented in sync with task-relevant goals, is found to be effective in inducing perceptual learning of nonspeech (Seitz et al., 2010) and non-native contrasts (Vlahou, Protopoulos, & Seitz, 2012).

This type of learning can be described computationally by reinforcement learning, whereby learning is driven by the outcome of feedback (e.g., reward or punishment) relative to a response. More specifically, learning emerges as one builds and updates predictions about the receipt of future reward (Sutton & Barto, 1998), thereby reducing the error signal in predicting reward (RPE) in subsequent trials. Incidental learning tasks like the video game may generate an RPE signal that propagates to multiple perceptual domains that support task success. In incidental learning tasks, learners have goals that are not directed to sound categorization but to other aspects of the task that incidentally promote sound category learning. Outcome is linked to task success, and learners may not be aware of the relationship between outcome and sound categorization. Therefore, the RPE signal generated during learning may modulate the representations of the auditory domain indirectly (see Lim et al., 2014).

In line with this possibility, language learning is supported by modulation from intrinsically generated attentional and motivational factors (Kuhl, Tsao, & Liu, 2003), contingent social cues (Goldstein, King, & West, 2003; Kuhl et al., 2003), and co-occurrence with multimodal information (Cunillera, Càmarra, Laine, & Rodríguez-Fornells, 2010; Cunillera, Laine, et al., 2010; Medina, Snedeker, Trueswell, & Gleitman, 2011; Roy & Pentland, 2002; Thiessen, 2010; Yeung & Werker, 2009). Even more directly, in a study investigating nonspeech auditory category learning in the same video game paradigm as the one used in the present experiment, Lim (2013) reports learning-dependent recruitment of the auditory regions supporting general auditory–phonemic category representations (i.e., left posterior STS; Desai, Liebenthal, Waldron, & Binder, 2008; Leech et al., 2009; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Liebenthal et al., 2010), as well as the striatum, implicated in RPE-based learning (e.g., Elliott, Frith, & Dolan, 1997; Poldrack et al., 2001; Schultz, Apicella, Scarnati, & Ljungberg, 1992; Tricomi, Delgado, McCandliss, McClelland, & Fiez, 2006; see Schultz, 2000, for a review). These results implicate reinforcement-related learning within incidental learning tasks and are consistent with the possibility that similar mechanisms supported the learning observed in the present study.

Conclusion

In the present work, we observed that participants can discover novel, acoustically variable functional units from continuous sound within an active video game task that does not involve overt categorization, segmentation or, in fact, explicit directed attention to the sounds at all. The sounds of the present experiments model the variability and regularity of language learning because they were derived from natural, continuous spoken language. Significantly, the learning observed exhibited the hallmark characteristic of category learning—generalization to novel stimuli. This generalization extended to isolated target words never experienced during training with continuous sound. The present experiments provide evidence that segmentation from continuous acoustic input without knowledge of the functional units, their temporal granularity, or the spectrotemporal acoustic dimensions relevant to defining them can occur in a relatively short time for the kinds of variability and regularity present in spoken language. Moreover, the results suggest the possibility that learning about the multiple levels of regularity that simultaneously unfold in natural spoken language may be supported by intrinsically generated learning signals evoked by more active tasks that include supportive multimodal associations.

References

- Ahissar, M., & Hochstein, S. (2002). The role of attention in learning simple visual tasks. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 253–272). Cambridge, MA: MIT Press.
- Altenberg, E. P. (2005). The perception of word boundaries in a second language. *Second Language Research*, 21, 325–358. <http://dx.doi.org/10.1191/0267658305sr250oa>
- Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.
- Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, 27, 387–414. <http://dx.doi.org/10.1017/S0272263105050175>
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403, 309–312. <http://dx.doi.org/10.1038/35002078>
- 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.
- Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109, 775–794. <http://dx.doi.org/10.1121/1.1332378>
- Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10, 512–528. <http://dx.doi.org/10.1093/cercor/10.5.512>
- Blesser, B. (1972). Speech perception under conditions of spectral transformation. I. Phonetic characteristics. *Journal of Speech and Hearing Research*, 15, 5–41. <http://dx.doi.org/10.1044/jshr.1501.05>
- 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. *The Journal of the Acoustical Society of America*, 101, 2299–2310. <http://dx.doi.org/10.1121/1.418276>

- Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*, 488–495. <http://dx.doi.org/10.3758/s13423-013-0501-5>
- Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates to Korean stops and fricatives. *Journal of Phonetics*, *30*, 193–228. <http://dx.doi.org/10.1006/jpho.2001.0153>
- Cole, R. A., & Jakimik, J. (1980). How are syllables used to recognize words? *The Journal of the Acoustical Society of America*, *67*, 965–970. <http://dx.doi.org/10.1121/1.383939>
- Colunga, E., & Smith, L. B. (2002). What makes a word? *Proceedings of the Annual Conference of the Cognitive Science Society*, *24*, 214–219.
- Cunillera, T., Càmarà, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*, 260–274. <http://dx.doi.org/10.1080/17470210902888809>
- Cunillera, T., Laine, M., Càmarà, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language*, *63*, 295–305. <http://dx.doi.org/10.1016/j.jml.2010.05.003>
- Cutler, A. (2000). Listening to a second language through the ears of a first. *Interpreting*, *5*, 1–23. <http://dx.doi.org/10.1075/intp.5.1.02cut>
- Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400. [http://dx.doi.org/10.1016/0749-596X\(86\)90033-1](http://dx.doi.org/10.1016/0749-596X(86)90033-1)
- Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*, 824–844. <http://dx.doi.org/10.1006/jmla.1994.1039>
- Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*, 1174–1188. <http://dx.doi.org/10.1162/jocn.2008.20081>
- Elliott, R., Frith, C. D., & Dolan, R. J. (1997). Differential neural response to positive and negative feedback in planning and guessing tasks. *Neuropsychologia*, *35*, 1395–1404. [http://dx.doi.org/10.1016/S0028-3932\(97\)00055-9](http://dx.doi.org/10.1016/S0028-3932(97)00055-9)
- Emberson, L. L., Liu, R., & Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization? *Cognition*, *128*, 82–102. <http://dx.doi.org/10.1016/j.cognition.2012.12.006>
- Fant, G. (1949). Analys av de svenska konsonantljuden. *L. M. Ericsson protokoll H/P* 1064.
- Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, *81*, 472–479. <http://dx.doi.org/10.1111/j.1467-8624.2009.01408.x>
- 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.
- Flege, J., & Wang, C. (1990). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t-/d/ contrast. *Journal of Phonetics*, *17*, 299–315.
- Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*, 3728–3740. <http://dx.doi.org/10.1121/1.418332>
- Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, *105*, 218–228. <http://dx.doi.org/10.1016/j.cognition.2006.09.005>
- Gauthier, I., & Tarr, M. J. (1997). Becoming a “Greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, *37*, 1673–1682. [http://dx.doi.org/10.1016/S0042-6989\(96\)00286-6](http://dx.doi.org/10.1016/S0042-6989(96)00286-6)
- 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*, 8030–8035. <http://dx.doi.org/10.1073/pnas.1332441100>
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54. <http://dx.doi.org/10.1016/j.cognition.2009.03.008>
- Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds “L” and “R.” *Neuropsychologia*, *9*, 317–323. [http://dx.doi.org/10.1016/0028-3932\(71\)90027-3](http://dx.doi.org/10.1016/0028-3932(71)90027-3)
- Grayson, J. H. (2006). Korean. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (2nd ed., pp. 236–238). Oxford, UK: Elsevier.
- Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, *25*, 577–588. <http://dx.doi.org/10.1037/0012-1649.25.4.577>
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*, 3099–3111. <http://dx.doi.org/10.1121/1.411872>
- Hockett, C. F. (1955). *A manual of phonology*. Baltimore, MD: Waverly Press.
- Holt, L. L. (2011). How perceptual and cognitive constraints affect learning of speech categories. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), *Handbook of laboratory phonology* (pp. 348–358). New York, NY: Oxford University Press.
- Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*, 1218–1227. <http://dx.doi.org/10.3758/APP.72.5.1218>
- Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate /r-/l/ on the basis of F3 onset frequency? *Bilingualism: Language and Cognition*, *15*, 434–435. <http://dx.doi.org/10.1017/S1366728912000041>
- 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. [http://dx.doi.org/10.1016/S0010-0277\(02\)00198-1](http://dx.doi.org/10.1016/S0010-0277(02)00198-1)
- Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*, 339–345. <http://dx.doi.org/10.1111/j.1467-7687.2009.00886.x>
- Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*, 701–714. <http://dx.doi.org/10.1121/1.406887>
- Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*, 323–328. [http://dx.doi.org/10.1016/S1364-6613\(99\)01363-7](http://dx.doi.org/10.1016/S1364-6613(99)01363-7)
- Kabak, B., & Idsardi, W. J. (2007). Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech*, *50*, 23–52. <http://dx.doi.org/10.1177/00238309070500010201>
- Klein, W. (1986). *Second language acquisition*. Cambridge, UK: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511815058>
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843. <http://dx.doi.org/10.1038/nrn1533>
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*, F13–F21. <http://dx.doi.org/10.1111/j.1467-7687.2006.00468.x>
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608. <http://dx.doi.org/10.1126/science.1736364>
- 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*, 9096–9101.

- Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *The Journal of Neuroscience*, *29*, 5234–5239. <http://dx.doi.org/10.1523/JNEUROSCI.5758-08.2009>
- Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*, 1621–1631. <http://dx.doi.org/10.1093/cercor/bhi040>
- Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., & Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, *20*, 2958–2970. <http://dx.doi.org/10.1093/cercor/bhq045>
- Lim, S.-J. (2013). *Investigating the neural basis of sound category learning within a naturalistic incidental task* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, Pennsylvania.
- Lim, S.-J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience*, *8*, 230. <http://dx.doi.org/10.3389/fnins.2014.00230>
- Lim, S.-J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*, 1390–1405. <http://dx.doi.org/10.1111/j.1551-6709.2011.01192.x>
- 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*, 683–698. <http://dx.doi.org/10.1162/jocn.2009.21392>
- 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. *The Journal of the Acoustical Society of America*, *94*, 1242–1255. <http://dx.doi.org/10.1121/1.408177>
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*, 874–886. <http://dx.doi.org/10.1121/1.1894649>
- Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.
- Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? In A. H.-J. Do, L. Dominguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 480–490). Somerville, MA: Cascadilla Press.
- Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 9014–9019. <http://dx.doi.org/10.1073/pnas.1105040108>
- Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*, 331–340. <http://dx.doi.org/10.3758/BF03211209>
- Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, *96*, 40–55. <http://dx.doi.org/10.1121/1.410492>
- Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*, 674–685. <http://dx.doi.org/10.1111/j.1467-8624.2009.01290.x>
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184. <http://dx.doi.org/10.1121/1.1906875>
- Pickett, J. M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, *6*, 151–164.
- Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*, 115–154. <http://dx.doi.org/10.1177/00238309030460020501>
- Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as “asymmetric sampling in time.” *Speech Communication*, *41*, 245–255. [http://dx.doi.org/10.1016/S0167-6393\(02\)00107-3](http://dx.doi.org/10.1016/S0167-6393(02)00107-3)
- Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. <http://dx.doi.org/10.1038/35107080>
- Pollack, I., & Pickett, J. M. (1964). Intelligibility of excerpts from fluent speech: Auditory vs. structural context. *Journal of Verbal Learning and Verbal Behavior*, *3*, 79–84. [http://dx.doi.org/10.1016/S0022-5371\(64\)80062-1](http://dx.doi.org/10.1016/S0022-5371(64)80062-1)
- Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*, 113–146. http://dx.doi.org/10.1207/s15516709cog2601_4
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. <http://dx.doi.org/10.1126/science.274.5294.1926>
- Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*, 199–207. <http://dx.doi.org/10.1038/35044563>
- Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *The Journal of Neuroscience*, *12*, 4595–4610.
- Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*, 2400–2406.
- Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9*, 565–573. <http://dx.doi.org/10.1111/j.1467-7687.2006.00534.x>
- Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*, 435–443. <http://dx.doi.org/10.1016/j.cognition.2010.03.004>
- Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, *49*, 2604–2610. <http://dx.doi.org/10.1016/j.visres.2009.08.003>
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568. <http://dx.doi.org/10.1016/j.cognition.2007.06.010>
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Thiessen, E. D. (2007). The effect of distributional information on children’s use of phonemic contrasts. *Journal of Memory and Language*, *56*, 16–34. <http://dx.doi.org/10.1016/j.jml.2006.07.002>
- Thiessen, E. D. (2010). Effects of visual information on adults’ and infants’ auditory statistical learning. *Cognitive Science*, *34*, 1093–1106. <http://dx.doi.org/10.1111/j.1551-6709.2010.01118.x>
- Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*, 53–71. http://dx.doi.org/10.1207/s15327078in0701_5
- Tricomi, E., Delgado, M. R., McClelland, 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. <http://dx.doi.org/10.1162/jocn.2006.18.6.1029>
- Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*, 363–381. <http://dx.doi.org/10.1037/a0025014>
- Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental*

- Science*, 10, 159–164. <http://dx.doi.org/10.1111/j.1467-7687.2007.00549.x>
- Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, 118, 2618–2633. <http://dx.doi.org/10.1121/1.2011156>
- Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, 118, 249–262. <http://dx.doi.org/10.1121/1.1923369>
- Watkins, A. J., & Makin, S. J. (2007). Steady-spectrum contexts and perceptual compensation for reverberation in speech identification. *The Journal of the Acoustical Society of America*, 121, 257–266. <http://dx.doi.org/10.1121/1.2387134>
- Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America*, 119, 597–607. <http://dx.doi.org/10.1121/1.2141003>
- Werker, J., & Tees, R. (1984). Cross-language speech perception evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63. [http://dx.doi.org/10.1016/S0163-6383\(84\)80022-3](http://dx.doi.org/10.1016/S0163-6383(84)80022-3)
- Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, 37, 278–286. <http://dx.doi.org/10.1037/h0080725>
- Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113, 234–243. <http://dx.doi.org/10.1016/j.cognition.2009.08.010>
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414–420. <http://dx.doi.org/10.1111/j.1467-9280.2007.01915.x>

Appendix

Variability of Target Words Across Sentence Contexts

In order to quantify the substantial acoustic variability across utterances of the target words, we segmented target words across all training and test sentence stimuli (40 stimuli/word) and measured acoustic features (pitch and formants) using Praat (version 5.3). Mean values were computed by averaging the acoustic measurements of the target word portion of each sentence. The acoustic variability of the target words is compared against the average acoustic values of 12 vowels spoken by 48 native English female speakers from Hillenbrand, Getty, Clark, and Wheeler (1995).

Table A1 Acoustic Measurements of English Target Words Used in Experiments 2 and 3

	Pitch (F0)			F1			F2		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum	Mean	Minimum	Maximum
Blue	268	201	333	383	322	494	1,368	1,263	1,492
White	224	181	286	545	445	671	1,767	1,476	1,978
Green	259	202	350	442	356	519	2,125	1,924	2,309
Red	223	166	281	539	465	616	1,780	1,348	1,977
Average female (Hillenbrand et al., 1995)	220	161	270	617	502	766	1,762	1,470	2,105

Received November 19, 2014

Revision received March 9, 2015

Accepted April 2, 2015 ■